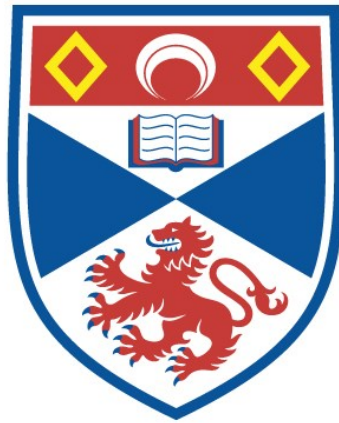


**CONTENT AND COMPUTATION : A CRITICAL  
STUDY OF SOME THEMES IN JERRY FODOR'S  
PHILOSOPHY OF MIND**

Mark Cain

A Thesis Submitted for the Degree of PhD  
at the  
University of St Andrews



1997

Full metadata for this item is available in  
St Andrews Research Repository  
at:

<http://research-repository.st-andrews.ac.uk/>

Please use this identifier to cite or link to this item:

<http://hdl.handle.net/10023/14711>

This item is protected by original copyright

# Content and Computation

*A Critical Study of Some Themes in Jerry Fodor's  
Philosophy of Mind*

Mark Cain

Ph.D. Thesis  
Department of Logic and Metaphysics  
University of St. Andrews



18 February 1997



ProQuest Number: 10167191

All rights reserved

INFORMATION TO ALL USERS

The quality of this reproduction is dependent upon the quality of the copy submitted.

In the unlikely event that the author did not send a complete manuscript and there are missing pages, these will be noted. Also, if material had to be removed, a note will indicate the deletion.



ProQuest 10167191

Published by ProQuest LLC (2017). Copyright of the Dissertation is held by the Author.

All rights reserved.

This work is protected against unauthorized copying under Title 17, United States Code  
Microform Edition © ProQuest LLC.

ProQuest LLC.  
789 East Eisenhower Parkway  
P.O. Box 1346  
Ann Arbor, MI 48106 – 1346

TL C 302

## Acknowledgements

Researching and writing this thesis has proved to be a long and arduous process during the course of which I have benefited from the help and encouragement of many people. I would like to take this opportunity to express my gratitude. First and foremost, I thank my parents for their unceasing support without which I would have ground to a halt several years ago. On the academic front, I have benefited from the supervision of Roger Squires, Barry Smith, and Sarah Patterson, and from discussions with Jim Edwards, Alan Millar, and, particularly, David Owens. Crispin Wright and Peter Clark provided some valuable practical advice and help. Many thanks to all of them. For their warm friendship and unstinting acceptance of my increasingly grouchy ways, my affection goes to Assunta del Priore, Paula Whiteside, Peter Dyke, David Owens, and Sarah Sawyer. Sarah Sawyer also proof read the penultimate draft and so any complaints about grammar, spelling, and punctuation should be addressed to her.

This thesis is dedicated to the memory of my friend Mark Powell and to the memory of my Grandmother Christiana Ashcroft Broadley.

# Abstract

In this thesis I address certain key issues in contemporary philosophy of mind and psychology via a study of Jerry Fodor's hugely important contributions to the discussion of those issues. The issues in question are: (i) the nature of scientific psychology; (ii) the individuation of psychological states for the purposes of scientific psychological explanation; and (iii) the project of naturalising mental content. I criticise many of Fodor's most significant and provocative claims but from within a framework of shared assumptions. I attempt to motivate and justify many of these shared assumptions.

Chapter 1 constitutes an overview of the key themes in Fodor's philosophy of mind. In Chapter 2 an account of scientific psychology within the orthodox computationalist tradition is developed according to which that discipline is concerned with explaining intentionally characterised cognitive capacities. Such explanations attribute both semantic and syntactic properties to subpersonal representational states and processes.

In Chapters 3 and 5 Fodor's various arguments for the conclusion that scientific psychology does (or should) individuate psychological states individualistically are criticised. I argue that there are pragmatic reasons why scientific psychology should sometimes attribute contents that are not locally supervenient. In Chapter 4 I consider Marr's theory of vision and conclude that the contents that Marr attributes to the states of the visual module are locally supervenient. Inconsistency is avoided by stressing the continuity of scientific psychological content with folk psychological content.

In Chapter 6 I develop an account of the project of naturalising mental content that vindicates that project. In Chapter 7 I address the question of whether Fodor's theory of content constitutes a successful engagement in that project. I argue for a negative answer before drawing some morals as to how we should proceed in the light of the failure of Fodor's theory.

## Declarations

I, Mark Cain, hereby certify that this thesis, which is approximately 100,000 words in length, has been written by me, that it is the record of work carried out by me and that it has not been submitted in any previous application for a higher degree.

date: 18 February 1997      signature of candidate:

I was admitted as a research student in October 1991 and as a candidate for the degree of Ph.D. in October 1991; the higher study for which this is a record was carried out in the University of St. Andrews between 1991 and 1997.

date: 18 February 1997      signature of candidate:

I hereby certify that the candidate has fulfilled the conditions of the Resolution and Regulations appropriate for the degree of Ph.D. in the University of St. Andrews and that the candidate is qualified to submit this thesis in application for that degree.

date: 18 February 1997      signature of supervisor:

In submitting this thesis to the University of St. Andrews I understand that I am giving permission for it to be made available for use in accordance with the regulations of the University Library for the time being in force, subject to any copyright vested in the work not being affected thereby. I also understand that the title and abstract will be published, and that a copy of the work may be made and supplied to any *bona fide* library or research worker.

date: 18 February 1997      signature of candidate:

# Contents

## Chapter 1

### **Introduction and Overview 1**

#### 1.1 Introduction

#### 1.2 An overview of Fodor's philosophy of mind

#### 1.3 An overview of the chapters to come

## Chapter 2

### **Psychological Explanation 15**

#### 2.1 Introduction

#### 2.2 The relationship to folk psychology

#### 2.3 Explaining cognitive capacities

#### 2.4 The nature of computation

#### 2.5 How could a physical system be a computer

#### 2.6 The role of meaning in psychological explanation

#### 2.7 Fodor's account of psychological explanation

## Chapter 3

### **Individualism and the Explanation of Cognitive Capacities 63**

#### 3.1 Introduction

#### 3.2 Individualism and folk psychology

#### 3.3 Individualism and computation

#### 3.4 Practical reasons for avoiding narrow content

#### 3.5 Capturing generalisations

#### 3.6 Explaining cognitive capacities

#### 3.7 Fodor's account of narrow content

#### 3.8 Conclusion

## Chapter 4

### **Individualism and Marr's Theory of Vision 63**

#### 4.1 Introduction

#### 4.2 Marr's theory of vision

#### 4.3 Burge's argument

#### 4.4 Circles and squares: a more graphic putative counter-example

#### 4.5 An externalist rejoinder

4.6 Behaviour versus environment

4.7 Conclusion

## **Chapter 5**

### **Causal Powers and a Metaphysical Argument for Individualism 141**

5.1 Introduction

5.2 The argument from causal powers

5.3 A putative counterexample

5.4 Crazy causal mechanisms and impossible laws

5.5 A modal argument

5.6 The nature of causal powers

5.7 Conclusion

## **Chapter 6**

### **The Naturalisation Project 180**

6.1 Introduction

6.2 The place of mental content in the natural world

6.3 Reductionism

6.4 Fodor's method

6.5 Normativity

6.6 Conclusion

## **Chapter 7**

### **Fodor's Theory of Content 209**

7.1 Introduction

7.2 Fodor's sufficient condition

7.3 Natural language

7.4 Laws and content

7.5 Asymmetric dependence

7.6 Some problem cases

7.7 Is there more to meaning than reference?

7.8 Conclusion

## **References 271**

# Chapter 1

## Introduction and Overview

### 1.1 Introduction

Jerry Fodor is one of the most important and influential philosophers of mind of the last thirty years or so. Since the mid-sixties he has been developing what might be described as a view, that is, a collection of distinct yet mutually reinforcing positions on a range of related issues. In this thesis my aim is to evaluate certain prominent elements of Fodor's view. These elements are: his account of the nature of scientific psychology and its relationship to folk psychology; his reflections on the question of whether scientific psychology is, or should be, individualistic; and his attempt to construct a naturalistic theory of content. In this introductory chapter I set myself the task of presenting an overview of Fodor's philosophy of mind along with an outline of the content of the chapters to come.

### 1.2 An overview of Fodor's philosophy of mind<sup>1</sup>

Throughout his career, Fodor's central aim has been to vindicate folk psychology within a broadly physicalist framework. This involves showing how the fundamental assumptions and theoretical commitments of folk psychology could be true given the physicalist theses that all real phenomena are identical to, or constituted by, physical phenomena, and that all properties that are instantiated in our world supervene upon the physical.

---

<sup>1</sup> In addition to his philosophical work, Fodor has made some important contributions to empirical psychology; in particular, to psycholinguistics and the study of mental architecture. See, for example, *The Psychology of Language* (co-written with Bever and Garret), and *The Modularity of Mind*. In this thesis I will largely ignore Fodor's psychological output, referring to it only when it bears directly on philosophical issues.



The details of Fodor's attempted vindication bear a profound imprint of certain key developments in the philosophy of mind and scientific psychology that took place during the early stages of his career. The developments in question are the overthrow of behaviourism and the cognitive revolution in scientific psychology, and the rise of functionalism - at the expense of both logical behaviourism and the type-identity theory - in the philosophy of mind.<sup>2</sup> These twin developments resulted in the establishment of a new orthodoxy in philosophy and scientific psychology according to which intentional mental states are physically realised internal states that cause behaviour and other intentional mental states, and whose essence (rather like states of a computer) resides at a level more abstract than the physical.<sup>3</sup>

As Fodor sees it, folk psychology is a descriptive, explanatory, and predictive practice that is bound up with a conception of the nature and workings of the human individual. According to this conception, human individuals stand apart from most of the other inhabitants of their world in being minded; that is, we are capable of thought and feeling and we behave or act rather than merely make movements. In this context, "thought" is a broad term that covers a wide range of distinct categories of mental states, including beliefs, desires, intentions, hopes, fears, expectations, and so on. In other words, thoughts are propositional attitudes (PA's for short), or intentional states. PA's are particularly important from the folk psychological perspective, for they play a prominent role in our mental life and in determining how we act or behave. This fact is reflected in the practice of folk psychology: much of folk psychological description, explanation, and prediction involves the

---

<sup>2</sup> Chomsky's critique of Skinner's behaviouristic account of language learning is a classic paper that both helped initiate the cognitive revolution in scientific psychology (and in the scientific study of the mind in general) and exerted a profound influence on Fodor. (See Chomsky (1959)). With respect to the rise of functionalism in the philosophy of mind, Hilary Putnam's work of the early 1960s stands out as being both hugely important and as having a particular influence on the development of Fodor's views. (See the papers collected in Putnam (1975a), in particular 'The Nature of Mental States' and 'The Mental Life of Some Machines').

<sup>3</sup> In fact, Fodor played an important role in the establishment of this new orthodoxy. See his early book *Psychological Explanation*.

attribution of PA's to human individual's and the employment of causal generalisations relating PA's to one another and to behaviour.

Given the centrality of PA's to the theory and practice of folk psychology, Fodor's task of vindicating folk psychology reduces to that of showing how we could have PA's; or rather, how we could have states that have the properties that folk psychology takes PA's to have.<sup>4</sup> The following are the key properties of, or the facts about, PA's that must be accounted for.

(i) PA's have semantic and intentional properties. For example, they have meaning or content, they are about things, they have satisfaction conditions, and so on. In general, having a PA involves standing in a relation to a propositional content. The *that* clauses of the sentences that we employ to attribute PA's to ourselves and our fellows serve to specify such propositional contents. For example, when I say that Edgar believes that Fang is ferocious I am saying that Edgar stands in the belief relation to the propositional content *Fang is ferocious*. (See 'Propositional Attitudes').

(ii) PA's are not causally inert; rather they are very much part of the causal fray. There are three main types of mental causation. First, PA's are often caused by environmental factors impinging on the subject, as when a display of Fang's ferocious behaviour causes me to believe that he is ferocious. Second, PA's often cause other PA's, as when my belief that Fang is ferocious causes a desire to avoid all contact with Fang. Third, PA's often cause behaviour, as when my belief that Fang is prowling nearby causes me to crouch behind the nearest bush. Consequently, folk psychological explanations that attempt to explain a behavioural episode, or the tokening of a PA, are causal explanations. The manner in which PA's causally interact with environmental impingements, other PA's, and behaviour is not random; rather it is regular or law governed. Consequently, there

---

<sup>4</sup> Fodor concentrates his attention on thought to the extent that he completely ignores feeling. In so doing, he accepts the idea that there is a fundamental distinction between cognition and consciousness, a distinction such that questions to do with the former can be addressed independently of questions to do with the latter. This idea is orthodox both within the cognitive science and the philosophical community. For a critique of such orthodoxy see Searle (1992). As a matter of fact, Fodor thinks that we have no hope of making any progress on the problem of consciousness at this point in our intellectual history.

are a whole battery of counterfactual-supporting causal generalisations relating PA's to one another, to environmental impingements, and to behaviour. The employment of such generalisations is fundamental to folk psychological description, explanation, and prediction. (See *Psychosemantics*, Ch. 1).

(iii) There is a systematic relationship between the causal powers of PA's and their semantic and intentional properties; PA's tend to cause PA's and behaviour to which they are semantically related. The most graphic case of this kind is constituted by demonstrative reasoning, a mental process where the contents of the thoughts that form the links of the process are related in a way that mirrors the relationship between the propositions of a logically valid argument. As Fodor puts it in *Psychosemantics*: 'one of the most striking facts about the cognitive mind as commonsense belief/desire psychology conceives it . . . [is] . . . the frequent similarity between trains of thought and *arguments*' (p. 13). This feature of mental processes is reflected in the nature of the generalisations employed by folk psychologists, many of which quantify over PA's and behaviour which are semantically or logically related. The classic example of such a generalisation is this: 'If X wants P, and X believes that not-P unless Q, and X believes that X can bring it about that Q, then, *ceteris paribus*, X tries to bring it about that Q' (*Psychosemantics*, p.13).

(iv) Thought is both productive and systematic. Each of us has the capacity to token any one of infinitely many content-distinct PA's. Moreover, anyone capable of having the thought that  $aRb$  (for any relation R, and subjects a and b) is also capable of thinking that  $bRa$ . For example, I believe that Fang is more ferocious than Edgar, and I am capable of believing that Edgar is more ferocious than Fang. (See 'Propositional Attitudes' and *Psychosemantics*, Appendix).

(v) Thought has the property of intensionality or opacity. In other words, PA contexts are not transparent to the substitution of co-referential expressions. For example, it can be true that Edgar believes that Mark Twain was born in Missouri, whilst false that he believes that Sam Clemens was born in Missouri, despite the fact that Mark Twain and Sam Clemens are one and the same. (See 'Propositional Attitudes' and 'Methodological Solipsism Considered as a Research Strategy in Cognitive Psychology').

Fodor notes that folk psychology is central to human life; we engage in folk psychology in all our dealings with our fellow humans. Consequently, if it turned out that folk psychology was false (in the sense that its basic categories had no reality, that it was based on a collection of false assumptions, and that it employed a body of generalisations that did not hold) we would suffer one of the greatest intellectual disasters imaginable. In fact, Fodor never seriously doubts folk psychology; for him, its everyday success is testament to its basic truth.<sup>5</sup> (See 'Three Cheers for Propositional Attitudes' and *Psychosemantics*, Ch. 1). Thus he attempts to construct a physicalist theory of mind that explains how folk psychology could be true; in other words, that explains how we could have mental states and processes that have the properties that are described in (i) - (v). His theory is a version of the Representational Theory of Mind (RTM for short) and can be described in the following terms.

PA's are relations to representations. That is, believing that Fang is ferocious involves standing in the belief relation to a representation that has the content *Fang is ferocious*. Representations are neurophysiologically realised, hence precisely the kind of items that can have causal powers. The belief relation (along with the desire relation, the hope relation, and so on) is to be understood in functional terms; whether or not a representation in my brain expresses the content of a belief depends upon how the mental processes that have access to that representation would manipulate it. Hence, PA's are like computational states in that they are relations to representations whose tokens are identical to, or constituted by, internal physical states, and yet are multiply realisable at the physical level.<sup>6</sup>

The representations that are the vehicles of PA contents belong to a language, namely the Language of Thought (LOT for short), or

---

<sup>5</sup> Thus he will have no truck with the eliminativism of Quine (1962), Feyerabend (1963), Rorty (1965), P.M. Churchland (1979, 1981), and P.S. Churchland (1986).

<sup>6</sup> Hence, Fodor rejects both the type-identity theory (see, for example, Place (1956), Feigl (1958), and Smart (1962)) and logical behaviourism (see, for example, Ryle (1949) and Hempel (1980)). For his explicit criticisms of the former see 'Special Sciences', and of the latter see 'Operationalism and Ordinary Language', (co-authored with Charles Chihara) *Psychological Explanation*, and *The Language of Thought*, Introduction.



Mentalese. LOT is not a natural language<sup>7</sup> but it is like a natural language in the respect that it comprises finitely many simple symbols and finitely many syntactic or grammatical rules for combining those simple symbols to create more complex symbol structures. Moreover, LOT has a combinatorial semantics in that the meaning of any complex symbol is determined by the meaning of its simple components and its syntactic structure. These similarities between LOT and a natural language such as English account for the productivity and systematicity of thought. Given that symbols of LOT can be combined to form longer strings of symbols, that in turn can be combined with other symbols to form yet longer strings, (and so on ad infinitum) there are infinitely many syntactically distinct sentences of LOT. Given that the meaning of a symbol string is determined by the meaning of its simple components and its syntactic structure, LOT will be capable of expressing infinitely many distinct meanings. And given that having a PA involves tokening a symbol of LOT, we will (at least in principle) be capable of tokening any of infinitely many content-distinct PA's. As for the systematicity of thought, if LOT is capable of expressing the proposition that *a* stands in relation *R* to *b*, then it will thereby be capable of expressing the proposition that *b* stands in relation *R* to *a*.

Fodor supplements RTM with a thesis about mental processes. Computers are systems which generate symbolic output from symbolic input by means of the application of symbol-manipulating rules. Computers are sensitive only to the formal or syntactic properties of the symbols that they manipulate, being blind to their semantic properties. However, they can be programmed in such a way as to respect semantic relations so that the symbols that they produce as output stand in appropriate semantic or logical relations to the symbols that they take as input. Fodor's idea is that the brain engages in computational activity, generating symbols of LOT from symbols of LOT by means of computation in such a way as to respect semantic and logical relations between the symbols so manipulated. Using Schiffer's metaphor of belief boxes, the idea is that what goes on when a belief that *P* interacts with a belief that if *P* then *Q* to cause

---

<sup>7</sup> There are philosophers committed to the existence of a language of thought who identify that language with the natural language of the thinker. See, for example, Harman (1972) and Carruthers (1996).

a belief that  $Q$  is as follows: a computational process in the subject's brain takes two symbols of LOT from the belief box (one of these being a symbol that has the content  $P$ , and the other being a symbol that has the content *if  $P$  then  $Q$* ), and generates as output a symbol of LOT that has the content  $Q$ , a symbol which it then places in the belief box. Hence, the idea that the brain engages in computational activity explains the fact that thought processes are generally logically and semantically coherent.

Fodor's version of RTM explains the intensionality of thought in the following manner. Whether or not a belief-ascribing sentence is true depends upon which sentences of LOT the subject has in his belief box. For example, for it to be true that Edgar believes that Mark Twain was born in Missouri, he has to have the LOT analogue of the English sentence "Mark Twain was born in Missouri" in his belief box. Clearly he can have that sentence in his belief box without having the LOT analogue of "Sam Clemens was born in Missouri" in his belief box. Hence it can be true that Edgar believes that Mark Twain was born in Missouri and false that he believes that Sam Clemens was born in Missouri, despite the fact that Twain and Clemens are one and the same. In general, it is due to the fact that having a PA involves tokening a sentence of LOT that a subject can think that  $a$  is  $F$  without thinking that  $b$  is  $F$  despite the fact that  $a=b$ .

At this point, it is worthwhile emphasising the fact that LOT is nothing like a private language in Wittgenstein's sense of the term.<sup>8</sup> There are three salient differences between LOT and a private language. First, the symbols of LOT do not refer to sensations. Second, the symbols of a private language are understood by the subject that employs them but cannot, even in principle, be understood by anyone else. With respect to LOT, matters are somewhat different. I do not grasp or understand the symbols of LOT that are manipulated in my brain. Neither do the computational processes that manipulate them, as they are blind to semantic properties. But cognitive scientists could, at least in principle, "crack the neural code" and so come to understand the sentences of LOT that are manipulated in my brain. Third, symbols of a private

---

<sup>8</sup> This is despite the fact that in *The Language of Thought* Fodor described LOT as a private language and attempted to defeat Wittgenstein's arguments for the conclusion that there could be no such language.

language get their meaning as a result of a conscious and deliberate act of ostensive definition. LOT, on the other hand, is innate, and the meanings of its symbols are not the product of any explicit naming ceremony.

What does Fodor offer by way of argument for his version of RTM? First, he thinks that it is the only theory that can account for the facts about PA's; it is "the only game in town". For example, he criticises traditional empiricism for being wedded to a theory of mental processes (namely, associationism) that cannot account for the semantic and logical coherence of thinking; he criticises logical behaviourism for having no account of thinking at all; and he criticises connectionism for its inability to account for the systematicity of thought.<sup>9</sup> Second, he argues that his theory has independent support from science; judged by purely scientific criteria, his theory is the best empirical hypothesis we have as to the nature and workings of the cognitive mind. This line of thought is pushed particularly hard in *The Language of Thought* where he argues that all currently plausible scientific psychological theories of decision making, concept learning, and perception, are committed to RTM. This reflects his official view that there is no fundamental divide between philosophy of mind and the scientific study of cognition: empirical investigation is relevant to philosophers of mind, as the questions that interest them cannot be answered by means of conceptual analysis or a priori reflection alone.

One problem with Fodor's theory as described thus far is that it doesn't account for the semantic and intentional properties of PA's. In virtue of what do the symbols of LOT have the semantic and intentional properties that Fodor attributes to them? Given his commitment to physicalism, Fodor is committed to the thesis that such properties are fixed or determined by non-intentional and non-semantic properties. But, ever since Brentano, it has been widely thought that intentional and semantic properties just couldn't be so fixed or determined and thus resist incorporation into the natural world. Hence there is a real challenge to physicalist friends of folk psychology (and intentional psychology in general). Fodor rises to this challenge by attempting to construct a naturalistic theory of

---

<sup>9</sup> See *Psychosemantics*, Appendix, and 'Connectionism and Cognitive Architecture: a Critical Analysis' (co-authored with Zenon Pylyshyn).

content, a theory that specifies the nonsemantic and nonintentional determinants of the semantic and intentional properties of the symbols of LOT. The bulk of this theory consists of a sufficient condition for a simple, nonlogical symbol of LOT to express a particular property.

Fodor's theory of content is essentially a version of informational semantics and is therefore atomistic. Fodor is vigorously opposed to holistic theories of content (theories according to which the meaning of a symbol depends on its relationships to other symbols) and devotes much energy to undermining the arguments for them. (See, in particular, 'Tom Swift and his Procedural Grandmother', *Psychosemantics*, Ch. 3, and *Holism: A Shopper's Guide* (co-authored with Ernest LePore)). His rabid opposition to holism is explained by the fact that he sees that doctrine as a threat to the viability of intentional psychology. His reasoning runs thus. There is no analytic-synthetic distinction. Consequently, if some of a symbol's relations to other symbols determine its meaning, then all of its relations to other symbols do. Given that distinct symbol-tokens rarely bear identical relations to other symbols, they rarely have the same meaning. In connection with LOT, barring a cosmic accident, no symbol of LOT in a subject's head is ever going to have the same content as that in the head of any of her fellows. The upshot of this is that distinct individuals rarely ever share the same PA's, and that intentional generalisations at most only ever subsume one individual. But if intentional generalisations only ever subsume one individual, then there are effectively no intentional generalisations; and if there are no such generalisations then intentional psychology is explanatorily and predictively toothless.

It is because of this threat that holism provides to the very viability of intentional psychology (and thus folk psychology) that Fodor will have no truck with functionalist theories of content (theories according to which the content of mental representations are determined by their causal roles). Hence, although Fodor was an early champion of functionalism and holds a functionalist account of PA relations, there are severe limits to the extent of his functionalism.

What, in Fodor's eyes, is the relationship between folk psychology and scientific psychology? Fodor's attempted vindication of folk



psychology implies that it comes close to being a bona fide special science as its states are physically constituted, its properties are physically determined, and its generalisations are hedged, true, counterfactual-supporting, and implemented by lower level mechanisms. Moreover, much of Fodor's reflections form part of a wider project of vindicating scientific intentional psychology. All this would tend to suggest that Fodor would hold that scientific psychology is little more than a rigorous, research-driven extension of folk psychology. In fact, he explicitly commits himself to such a view (see *Psychosemantics*, Ch. 1). However, for several years he held that there is a fundamental difference between folk and scientific psychology, as the latter individuates psychological states in such a way as to respect the local supervenience of the psychological on the physical, and thus employs a notion of narrow content. In other words, due to the fact that folk psychological content is broad (as is indicated by the familiar Putnam and Burge thought experiments), there is a mismatch between the respective taxonomies of the two psychologies. This placed on Fodor the additional burden of giving a plausible account of narrow content. He did not shirk from this task, describing narrow content as a function from contexts to broad contents.

More recently, Fodor has abandoned narrow content as he has become reconciled to the idea that an intentional psychology would not need to go narrow in order to satisfy the rigorous demands of scientific methodology.

That completes my description of Fodor's philosophy of mind. In short: Fodor engages in the project of vindicating folk psychology within a physicalist framework, and doing so in such a way as to establish that it constitutes a firm basis for the development of a respectable scientific intentional psychology. Central to that vindication is (i) a version of the Representational Theory of Mind, according to which PA's are computational relations to symbols of a Language of Thought, and thought processes are computational processes involving the manipulation of such symbols; and (ii) an informational theory of content that specifies the nonsemantic and nonintentional determinants of the contents of the symbols of LOT and, thereby, of our PA's.

### 1.3 An overview of the chapters to come

This thesis does not constitute a comprehensive account and study of Fodor's philosophy of mind, as several key elements of his philosophical output are, in effect, ignored.<sup>10</sup> Rather, what I attempt to do is address certain key issues in contemporary philosophy of mind and psychology via a study of Fodor's contributions to the discussion of those issues. The issues in question are as follows. (i) The nature of scientific psychology. How does scientific psychology relate to folk psychology? What are its explanatory ambitions and basic theoretical assumptions and commitments? What are the respective roles of semantic and syntactic properties in its explanations? (ii) The individuation of scientific psychological states. Does (or should) scientific psychology individuate the states that figure in its explanations individualistically (for example, in terms of their narrow content)? Do the basic commitments, assumptions, and explanatory aims of scientific psychology tell either for or against individualism? Are there any relevant metaphysical considerations? (iii) The project of naturalising content. What is the naturalisation project? Is it a misguided project or does the scientific status of psychology depend upon the possibility of naturalising content? Does Fodor succeed in naturalising content? If not, what morals are to be drawn from the failure of his theory?

As will become apparent, there is much that I admire in Fodor's approach to these issues. Moreover, I endorse, or at least sympathise with, many of his basic assumptions. I accept Fodor's account of folk psychology; it is a proto scientific theory of the nature and workings of the human individual that permeates our dealings with our fellows. I also share Fodor's intentional realism; we really have beliefs, desires, and the like and the generalisations of folk psychology are largely true. Further, folk psychology is closely related to scientific psychology and does constitute a sound basis for the construction of such a psychology. I accept Fodor's physicalism; the sum totality of facts is ultimately determined by the sum totality of physical facts and science is, by its very nature, committed to such a doctrine. Finally, I agree with Fodor that empirical research and

---

<sup>10</sup> For example, his attempts to undermine the arguments for holism, and his attack on connectionism.

findings are relevant to what have traditionally been seen of as philosophical questions that can only be answered by means of a priori reflection or conceptual analysis. As a result of all this, I view Fodor's project of vindicating intentional psychology (both in its folk and scientific varieties) within a broadly physicalist framework as being a viable and important project. However, I wish to dispute many of Fodor's specific pronouncements on the issues that are my concern. Thus I should be seen not as a disciple of Fodor but as a sympathetic and admiring detractor and critic.

The structure of my thesis can be described in the following terms. In Chapter 2 I develop an account of the nature of scientific psychology. According to this account, the aim of scientific psychology is to account for our intentionally characterised cognitive capacities. It does this by descending to the sub-personal level and so posits a whole range of representational states and processes that are unfamiliar from the folk psychological perspective. Scientific psychological explanations of cognitive capacities do, and must, appeal to both semantic and syntactic properties of representations and representation-manipulating processes.

Chapters 3, 4, and 5 address the question of whether scientific psychology is (or should be) individualistic, or should employ a notion of narrow content. Here my reflections rely heavily on the claims made in Chapter 2. In Chapter 3 I consider those of Fodor's pro-individualistic arguments that appeal to psychological assumptions and practice. These arguments fail as they rely on a misrepresentation of those assumptions and that practice. There are good practical reasons why scientific psychology should avoid "going narrow" if at all possible, especially if narrow content is to be understood as Fodor describes it. I argue that these practical considerations may well carry the day given the contingent fact that we have yet to discover any other worldly twins and cousins. Thus, I make a tentative externalist conclusion: some of the intentional properties that scientific psychology attributes to our representational states in the course of explaining our cognitive capacities are not locally supervenient. And that this is so is all to the good.

In Chapter 4 I turn away from Fodor to consider Marr's theory of vision; is that theory individualistic? I argue against a negative conclusion; the contents that Marr attributes to the states of the

visual module are locally supervenient. I then attempt to reconcile this conclusion with that of Chapter 3 by arguing that the contents that Marr attributes find echo in those that folk psychology attributes to our personal level visual states and experiences as those contents are locally supervenient. Thus the general conclusion of Chapters 3 and 4 is that some (but not all) of the contents that scientific psychology does (and should) attribute to our psychological states are not locally supervenient. In Chapter 5 I attempt to defend this conclusion against Fodor's metaphysical arguments for individualism. These arguments rest upon the idea that science individuates in terms of causal powers and that the causal powers of psychological states are locally supervenient. I argue that the causal powers of psychological states are not locally supervenient by developing an interest-relative account of causal powers.

Chapters 6 and 7 are concerned with the naturalisation project. In Chapter 6 I develop an account of that project that reveals it to be both sensible and important. Making progress in the project of naturalising content is a central component of the program of vindicating scientific psychology in such a way as to discredit its ever-vocal Brentano-inspired detractors. Naturalising content does not involve reducing content to something non-semantic and non-intentional; rather, it involves generating naturalistic sufficient conditions that specify the non-semantic and non-intentional determinants of the semantic and intentional properties that scientific psychology attributes to our states.

In Chapter 7 I address the question of whether Fodor's theory is a successful engagement in the naturalisation project as I have described it. I argue for a negative conclusion. Once again I rely upon the account of scientific psychology developed in Chapter 2 by appealing to representational states that figure in its explanations that are quite other than the familiar everyday beliefs that populate Fodor's imagination. A consideration of such states raises doubts as to the generality of Fodor's theory, and generates counterexamples to it.

Before closing this chapter I will outline some of the prominent themes that emerge in the course of my reflections. First, when addressing the issues of individuation and naturalisation it is of crucial importance to bear in mind the explanatory ambitions of

scientific psychology and its basic theoretical assumptions and commitments. Second, it is important not to focus on too narrow a range of examples. Fodor makes just this mistake. For example, when he engages in the naturalisation project he has a tendency to focus his attention on a small number of cases that seem familiar and salient from the folk psychological perspective. As a result of this, certain of his claims have an air of plausibility that they would lose if only we cast our eye more widely. And when he argues for individualism he focuses almost exclusively on the case of twins with beliefs featuring natural kind concepts. This tends to obscure the fact that narrow psychology (at least as Fodor conceives it) is not a practical option if our aim is that such a psychology should apply to our cousins as well as our twins. Third, it is important not to overlook practical considerations and contingencies. There is a danger of making demands of scientific psychology that are so great that its practitioners have no real hope of satisfying them. However, a less than pure psychology that we have a chance of making some progress in is much to be preferred to a pure psychology that we don't have the ability to engage in. In particular, I have in mind the issue of individuation. Suppose that an ideal scientific psychology would be narrow in virtue of the generality of such a psychology. It wouldn't follow from that fact alone that a respectable scientific psychology should be narrow. For it may well be that the practical difficulties of engaging in narrow psychology outweigh the advantages of greater generality. This will especially be the case if, as a matter of fact, we have no twins or cousins who are similar to us in their narrow psychological states.

It is now time to get down to some work.



## Chapter 2

# Psychological Explanation

### 2.1 Introduction

In this chapter my central aim is to develop an account of the nature of scientific psychology and the explanations that it endeavours to produce. This account will underlie my discussion of individualism and the naturalisation project in subsequent chapters and will motivate many of my criticisms of Fodor's most interesting and provocative claims.

I make no claims to present an account that applies to everything that can justifiably be labelled "scientific psychology". The science of psychology, like most sciences, is a broad discipline consisting of subdivisions that vary widely in terms of their explanatory aims, research methods and theoretical assumptions. Like many contemporary philosophers of mind, when I talk about scientific psychology what I really mean is cognitive psychology, that branch of psychology that seeks to explain our cognitive capacities. However, my account doesn't even apply to everything that can justifiably be described as cognitive psychology. For example, connectionist theories<sup>11</sup> and J.J. Gibson's theory of vision<sup>12</sup>, though both arguably cognitive psychological theories, certainly do not fit my picture. What I intend my account to apply to is most mainstream cognitive psychology of the last three decades or so. Such psychology has been variously described as orthodox computationalism (Cummins, 1989), as being committed to the Computational Theory of Mind (Fodor, 1980), and as viewing the mind as a physical symbol system (Newell and Simon, 1976; Newell, 1980). As these characterisations suggest, the idea that in some important respect the mind is a computer is a defining assumption of this brand of psychology. Another central characteristic of it is that it is an intentional psychology; that is,

---

<sup>11</sup> See, for example, Rumelhart and McClelland (1986).

<sup>12</sup> See, for example, Gibson (1979).

intentional states figure prominently in its theories and explanations.

A satisfactory account of mainstream cognitive psychology<sup>13</sup> should specify the following:

- (i) The explanatory ambitions of scientific psychology.
- (ii) Its basic ontological commitments and theoretical assumptions; in particular it should spell out the nature of the Computational Theory of Mind that underlies it.
- (iii) Its relationship to folk psychology, that is, the descriptive, explanatory and predictive practice that everyday folk engage in in their dealings with their fellows.
- (iv) Its relationship to lower level sciences such as neuroscience and physics.
- (v) The role of semantic and intentional properties in its explanations.

Hopefully my account will satisfy all these requirements.

## **2.2 The relationship to folk psychology**

In order to describe the relationship that scientific psychology bears to folk psychology I shall begin with a brief account of some salient features of the latter. Folk psychology is bound up with a conception of the human individual that can be described in the following terms. We stand apart from most of the entities that exist in our world by being minded. Related to our being minded is the fact that we don't just make movements or have things happen to us; rather, we behave or act. Part of what it is to be minded is to have or experience mental states of which there are many distinct types. One broad category of mental state types contains states that have intentional or semantic properties, properties in terms of which they are identified and individuated. These are the propositional attitudes, the most familiar being beliefs, desires and intentions. Propositional attitudes (PA's for short) are not causally inert. In addition to being caused by environmental impingements on the

---

<sup>13</sup> From here on in I will use the term "scientific psychology" in place of the more cumbersome "mainstream cognitive psychology".

sensory apparatus, they cause, and interact with one another to cause, other PA's, non-intentional mental states and behaviour. These causal episodes are not random; on the contrary, there are countless counterfactual-supporting causal generalisations in which PA's figure. The causal powers of PA's are such that, as a general rule, they cause PA's and behaviour to which they are logically or semantically related, as when a desire for a bottle of beer and a belief that there is one in the fridge interact to cause an intention to get that bottle, an intention that subsequently causes the subject to go to the fridge and retrieve the bottle. Consequently, the causal chains in which PA's figure are typically rationally coherent, and this fact is reflected in the nature of the generalisations in which they figure. Finally, in virtue of the manner in which environmental impingements, PA's, other mental states, and behaviour interact with one another, we all have a battery of cognitive capacities including those of being able to acquire knowledge of the external world by means of perception, recognise faces, remember past events, categorise objects, solve problems, and so on.

However, folk psychology isn't merely an abstract theory. Rather it is a descriptive, explanatory and predictive practice that is central to human life. It plays a fundamental role in all our dealings with our fellow humans; without folk psychology we would be at sea in the social world. In describing, predicting and explaining mental and behavioural episodes we employ the above described causal generalisations. Given the nature of these generalisations, a typical folk psychological explanation of a PA or a behavioural episode will represent it as being caused by something to which it is semantically or logically related. Hence, such explanations don't just indicate the causes of our PA's and behaviour; in addition they make sense of them or reveal them to be rational.<sup>14</sup> And because we believe that people are by and large rational we only rest satisfied with a candidate

---

<sup>14</sup> To say that I am scattering salt on my lawn because I want to get rid of the dandelions growing on it and believe that salt kills dandelions is to specify the causes of my behaviour. And it is also to make sense of my behaviour, to reveal it to be rational in the light of its causes. To explain my behaviour by an appeal to the belief that Kilimanjaro has snow-covered peaks and the desire for Fang to lose his teeth is not to so make sense of my behaviour due to the absence of any sensible semantic or logical relation between the causes and their effect.



causal explanation when it makes sense of the effect in question, when it reveals it to be rational or sensible in the light of its causes and their doxastic surroundings.<sup>15</sup>

Now how does scientific psychology relate to folk psychology? Is the former merely a sharpened up, more rigorous version of the latter? Or is it bound up with a distinct body of theoretical assumptions and concerned with a different explanatory project?

It would perhaps be a surprise and out of keeping with the general nature of the relationship between folk theories and their scientific counterparts were it to be the case that folk and scientific psychology attempted to explain the same range of phenomena. What we should expect is for them to have somewhat different explanatory agendas. The point of folk theories is to enable us to deal and interact successfully with our environment on an everyday basis. In order to do this we need to be able to describe and categorise phenomena accurately and quickly and, on the basis of such descriptions, construct explanations and predictions of particular events. We behave in the light of such descriptions, explanations and predictions, and thus our success depends on their truth value; for example, if we were generally way out in our folk physical descriptions, explanations and predictions, we could hardly prosper in a world of medium-sized physical objects. Given that we utilise generalisations in the construction of such descriptions, explanations and predictions, folk generalisations must, by and large, approximate the truth. An important point is that ordinary folk, though they utilise such generalisations, are rarely concerned with the question of why these generalisations hold. For example, most of us know that unsupported objects fall to the ground, and that water expands when frozen. It is important for us to know such facts, but for ordinary, everyday purposes it just doesn't matter why they hold. For the purposes of explaining such events as the cracking of Edgar's pipes on a cold winter's night and avoiding such misfortunes myself, I will need to know that water expands when frozen, and that water in pipes tends to freeze on cold winter nights. But as to why water behaves in this way I need know nothing; it is only outside the folk

---

<sup>15</sup> For an extended account and celebration of folk psychology very much in the spirit of my pronouncements see *Psychosemantics* Ch. 1.

physical arena that such knowledge will be of any use or interest to me.

All this suggests the following. The folk practice of explaining and predicting is typically one of explaining and predicting particular events. The generalisations that are employed in such activity are taken for granted; it is rarely a folk enterprise to explain why they hold. In the case of folk psychology, the concern is to explain and predict particular behavioural and mental events but not to account for the generalisations or facts about the mind the recognition of which is central to such explanation and prediction. Science, on the other hand, is somewhat different. Scientists don't just accept the facts or generalisations recognised by ordinary folk; rather they attempt to explain why they hold. And often, in constructing such explanations, they postulate a whole range of phenomena and generalisations not recognised by ordinary folk. Thus it is a job for physics to explain why water expands when frozen, and the explanation will involve reference to molecules, molecular bonds, the effect of temperature-falls on them, and so on. Similarly, folk biology recognises that individuals inherit certain properties of their parents but it takes scientific biology to account for such facts, and in doing so refers to a whole range of properties and phenomena not recognised by its folk counterpart, namely, genes, chromosomes, DNA, and the like.<sup>16</sup>

---

<sup>16</sup> Expressed in the terminology of Cummins (1983) this comes pretty close to the claim that folk explanations are transition theories whereas science aims at property theories. Cummins outlines the distinction in these terms:

The point of what I call a *transition theory* is to explain changes of state in a system as effects of previous causes - typically disturbances in the system. The emphasis is on what will happen *when* (i.e. under what conditions). Subsumption under causal law is the natural strategy: one tries to fix on a set of variables for the system that will enable one to exhibit each change of state as a function of a disturbing event and the state of the system at the time of the disturbance. (pp. 2-3)

The point of what I call a *property theory* is to explain the properties of a system not in the sense in which this means "Why did S acquire P?" or what caused S to acquire P?" but, rather, "What is it for S to instantiate P?", or, "In

This way of describing matters makes it look as if there is a respect in which a science's agenda is set by its folk relative and that sciences generally endorse the ontological commitments and accept the generalisations of their folk counterparts. But, of course, this is not true without qualification. Often scientists attempt to explain facts and generalisations not recognised by any folk theory, sometimes they question the reality of folk categories, and sometimes they conclude that folk seriously misrepresent the facts. But all this notwithstanding, there is usually a close relationship between folk theories and their scientific counterparts: the latter are born of the former and rarely serve to completely undermine their parent. Another feature of folk theories is that they are frequently enriched by developments in science; ordinary folk often talk of atoms, molecules, gravity, genes, unconscious desires, long term memory, and the like. Thus I am idealising somewhat when I say that science seeks to account for the generalisations invoked, or facts assumed, in the construction of folk explanations, and that in doing so it appeals to a whole range of properties (and generalisations quantifying over them) that are not recognised by ordinary folk. But it is of the nature of idealisations that they approximate rather than fundamentally misrepresent the truth.

Hence we should expect scientific psychology to have a different explanatory agenda to folk psychology, to appeal to a whole range of properties, phenomena and generalisations not recognised by it, and

---

virtue of what does S have P?" . . . The natural strategy for answering such a question is to construct an analysis of S that explains S's possession of P by appeal to the properties of S's components and their mode of organisation. The process often has as a preliminary stage an analysis of P itself into properties of S or S's components. (p. 15)

I accept that there is a distinction between explaining a particular event and explaining a fact about a system (or a generalisation true of it) that the explanation of that event presupposes. And I accept that science typically aims at the latter sort of explanation, explanations that often satisfy Cummins' description of a property theory. However, as will become apparent later, I am not at all convinced that the distinction between transition theories and property theories is as clear-cut as Cummins would have us believe. Hence I am reluctant to say that science in general, and scientific psychology in particular, aims to construct property theories.

yet to share many of its most basic commitments and assumptions. Even the most cursory examination of folk and scientific psychology and the nature of the relationship between the two bears out this expectation. The primary concern of folk psychological explanation is to explain particular events, namely instances of behaviour and tokenings of mental states. It is not a folk psychological concern to explain why the generalisations employed in the construction of such explanations hold; rather they are just assumed to hold. It is a fundamental assumption of folk psychology that we have a whole battery of cognitive capacities; having such capacities is part of what it is to be minded. Many folk psychological generalisations correspond to such capacities; for example, the generalisations that people understand sentences of their own language whenever they are presented with such a sentence, and that people form true beliefs about the nature of the world on the basis of their perceptual experiences, correspond, respectively, to the cognitive capacities to understand sentences of one's own language and to perceive and classify objects in one's immediate environment. Such capacities are intentionally characterised, for exercising them typically involves tokening an intentional state, a belief, for example.

Folk psychology offers little by way of explanation of our cognitive capacities; it doesn't answer such questions as how we perceive objects, how we understand spoken and written sentences, how we recognise faces, how we recall past events, and so on. This is where scientific psychology comes in, for it attempts to answer such "how" questions. Thus it shares with folk psychology the assumption that we have cognitive capacities (and many of the cognitive capacities recognised by folk psychology) and that we have the intentional states (beliefs, for example) that are tokened in the exercise of these capacities. Yet despite being committed to the reality of beliefs and other PA's - and causal generalisations in which they figure - scientific psychology is not a belief-desire psychology, for in accounting for our cognitive capacities it appeals to a quite different range of states and processes. The states in question are subpersonal representational states. They are states that play a role in the proximal causation of beliefs, perceptual experiences, and so on, without belonging to any such category of personal level state; they are, in Stich's (1978) terminology, subdoxastic states. And the



processes in question are subpersonal representation manipulating processes.

How are we to understand the distinction between personal level intentional states and processes and subpersonal representational states and processes?<sup>17</sup> The distinction is difficult to characterise precisely, but at an intuitive level it is easy enough to grasp. Personal level states, such as beliefs, are states of a whole person not of a person's parts. It is I, a whole person, who believes that Fang is ferocious. None of my proper parts has this belief or any other belief. Hence belief attributions are attributions of an intentional state to a whole person; it would be a confusion to attribute a belief to a system and deny that it was a person. Similarly, certain mental processes are executed by whole persons and not by any of their proper parts; for example, it is me, not any of my parts, who engages in mental arithmetic when calculating my bank balance, and who reflects on the writings of Jerry Fodor. Subpersonal states, on the other hand, are not states of a whole person but, rather, states of a person's proper parts. For example, the primal sketch (which represents significant changes in light intensity across the retinal image, (see Marr, 1982)) is a state of a component of the brain, namely the visual module. When a primal sketch is tokened within me I do not represent a pattern of changes in light intensity across a retinal image, rather a part of my brain does that. Similarly, some mental representational processes are executed not by whole persons but by their parts. For example, stereopsis, the process of extracting depth information from disparities between a pair of retinal images, is executed by the visual system (or a component of it). Such processes are subpersonal.

Being constituted by an internal physical state is not enough to make a representational state subpersonal. Hence, even if beliefs are constituted by brain states, it doesn't follow that they are subpersonal states, even though brain states are states of a part of a person. If that sounds odd consider the following. Some medical conditions are constituted by internal physical states. Edgar's arthritis is constituted

---

<sup>17</sup> The personal-subpersonal distinction was originally made by Dennett (1969). According to his characterisation, explanation at the personal level is intentional and instrumental, whereas subpersonal explanation is mechanical and physiological. In later writings (see his 1981, for example) he describes subpersonal states as having informational content.

by a state of his joints. The constituting state is thus a state of one (or several) of his parts, namely his joints. Nevertheless it is Edgar the whole person, as opposed to his joints, that has arthritis. Joints don't get arthritis; it is people that suffer that misfortune. Here is another example, this time involving a process. As I run away from a snarling Fang I break into a sweat. My sweating is constituted by a physiological process that takes place at the surface of my skin. Yet it is I that sweat and not my skin. My claim then is that if beliefs and such mental processes as thinking about Jerry Fodor's writings are constituted by brain states and processes they will not thereby be subpersonal, rather they will be personal level phenomena analogous to arthritis and sweating in the above described respect.<sup>18</sup>

In short, then, we can characterise the relationship between folk and scientific psychology in the following terms: scientific psychology endorses many of the basic assumptions and ontological commitments of folk psychology, but whereas the latter attempts to causally explain and predict particular behavioural and mental events, and in doing so sticks to the personal level, the former attempts to explain our cognitive capacities by descending to the subpersonal level.

---

<sup>18</sup> I accept that this characterisation of the personal-subpersonal distinction could do with being somewhat more rigorous and precise. However, I think I have succeeded in giving the reader a basic feel for what is arguably a very familiar and obvious distinction. I suspect that to attempt to do anything more than this would be to engage in a lengthy and arduous project that would take me far off track.

Perhaps I ought to mention that I don't think there is much hope of grounding the distinction between personal level and subpersonal representational states in either that between conscious and unconscious states or that between states with conceptual content and states with nonconceptual content. An appeal to conscious awareness won't work because of the existence of unconscious personal level states of the sort that populate psychoanalytic theories. Moreover, to echo a point made by Davies (1989), we could conceive of a subpersonal representational state that surfaced in conscious awareness as a distinctive type of sensation. An appeal to the distinction between conceptual and nonconceptual content fares no better as some personal level representational states have nonconceptual content (visual experiences, for example) and, arguably, some subpersonal states have conceptual content (for example some of those involved in language comprehension and production) (Peacocke, 1992).

### 2.3 Explaining cognitive capacities

What specific form do scientific psychological explanations of our cognitive capacities take? What does the scientific psychologist offer by way of answers to such "how" questions as the following: how do we acquire knowledge about the nature of the external world by means of vision?; how do we understand sentences?; how do we recognise faces?; how do we categorise objects?; how do we remember past events? So far I have said little more than that the scientific psychologist descends to the subpersonal level presenting explanations that appeal to subpersonal representational states and processes. In itself this is hardly very illuminating. A first valuable step in remedying this situation involves reflecting on the question of how capacities in general are explained.

Many capacities - not just cognitive capacities - are complex in the sense that to have them a system (or its parts) must have a whole series of simpler capacities. The complex capacity is executed by executing these simpler capacities in a certain order. A familiar example comes from cooking: in order to have the capacity to bake a cake one must have such simpler capacities as those of being able to break an egg, weigh flour, sugar and the like, mix flour, eggs and sugar together in a bowl, and so on. In executing the capacity to bake a cake one has to execute these simpler capacities in a certain order. One answers the question "how do you bake a cake" by specifying an ordering of tasks simpler than that of baking a cake such that the execution of those tasks in that order reliably lands one with a baked cake. Psychology attempts to account for our cognitive capacities in much the same way; that is, by specifying the simpler capacities that our cognitive capacities depend upon, and describing the order in which they are executed whenever the cognitive capacity in question is exercised.

Cognitive capacities are intentionally characterised; their exercise involves the tokening of a personal level intentional state and they could not be had by a system that did not token, or was not capable of tokening, such states. These capacities are close to being miraculous. For example, light reflected off objects in my local environment hits my retina and within a fraction of a second I have a rich perceptual experience in which the world is presented to me as being a certain

way (a way pretty much how it is) and I form a whole collection of true beliefs about the nature of my local environment. How could we be capable of such feats?

A fundamental assumption of scientific psychology is that the simpler capacities underlying our cognitive capacities are information processing capacities of our brain and its parts. Parts of the human brain have the capacity to do such things as the following: generate information from information, and in so doing extract, or make explicit, new information; compare separate items of information; and store and retrieve information when appropriately prompted.<sup>19</sup> In processing information the brain doesn't generate, store and retrieve any old information but, rather, information that is relevant to our cognitive capacities and thus capable of supporting them. Thus, whenever we exercise a cognitive capacity our brain engages in information processing activity and in so doing generates information that is relevant to the task in hand, information that somehow breaks through to, or makes its mark on, the personal level.

But how is the brain able to process information? Attributing information processing capacities to the brain and its parts gives rise to two related worries. First, these capacities appear to be so similar to the cognitive capacities that they are invoked to explain that to appeal to them is to make no real explanatory progress. Second, these capacities appear to be the sort of capacities that the brain or its parts couldn't have short of being intelligent agents, something that they are not. This brings us to the second fundamental assumption of scientific psychology, an assumption that goes some way towards alleviating these worries. This is the assumption that it is by means of computation that the brain and its parts process information, computation being a process of mechanical symbol-manipulation. So, the idea is that the brain performs computational operations on physically realised symbols or representations, and in so doing processes information.

The second assumption has bearings on the nature of psychological explanation for it entails that the explanation of a cognitive capacity

---

<sup>19</sup> I am using the term "information" in its ordinary, everyday sense and not as a technical term. When I say that a symbol or a state carries or encodes information all I mean is that it represents something in the world as being a certain way.



must specify the computational means by which the brain and its parts exercise their information processing capacities. But this raises a whole series of questions. What exactly is computation? How could the brain, or any of its parts, manage to be a computer or have computational capacities? Just what role does a specification of the computational capacities and activity of the brain, and its parts, have in an explanation of a cognitive capacity? These questions need to be answered in the course of the construction of a complete account of the nature of scientific psychological explanation. But before I attempt to address them an important point needs to be made.

A complete explanation of a particular capacity of a system involves specifying a range of simpler capacities of the system (and/or its parts) such that their execution in the indicated order reliably results in an exercise of the target capacity. The explanations of contemporary scientific psychology fail to be complete in this respect. The information processing and computational capacities and activity that scientific psychologists describe resides at the subpersonal level. Consequently, what is required for their explanations to be complete is an account of precisely how activity at the subpersonal level generates personal level intentional states, an account, in other words, of the precise relationship between the subpersonal and the personal. It is my contention that contemporary scientific psychology has no such account, due as much as anything to its not having a complete account of the nature of personal level intentional states nor an account of how those states are realised in us. Consider Marr's theory of vision, for example. Marr's explicit aim is to account for our ability to acquire knowledge about the nature of the external world by means of vision. Yet he doesn't tell us how we reliably acquire true beliefs about the nature of the external world by means of vision but only how the visual module reliably generates, from a pair of retinal images, accurate, object-centred 3-D representations that indicate the shape, size, colour, texture etc. of distal stimuli. His assumption is that the generation of such representations plays an important role in belief-fixation, but he doesn't specify how they result in the formation of a belief (or, indeed, how they are connected to visual experiences). In other words, there is a gap in Marr's theory, a gap between the personal and the subpersonal. This gap is a quite

general feature of scientific psychological explanations of our cognitive capacities.<sup>20</sup>

The above point has bearings on one of Fodor's central arguments for RTM as a theory about the familiar PA's of folk psychology. As we saw in Chapter 1, Fodor argues that RTM underlies virtually all contemporary scientific psychology. This, he thinks, gives us good reason to endorse the theory. However, the RTM that scientific psychology endorses is not a theory about beliefs, desires, and the like, but rather a theory about subpersonal representational states. The representations that populate scientific psychological theories, the primal sketch and the 3-D representation, for example, are not the vehicles of belief-contents. Thus Fodor has misrepresented the nature of the RTM that scientific psychology is committed to.<sup>21</sup>

---

<sup>20</sup> It might be thought unacceptable that there is such a gap in scientific psychological explanations; that, contrary to the advertisements, they don't account for any of our cognitive capacities. But such a judgement would surely be unacceptably harsh. Minds and their capacities are so very complex and scientific psychology is such a young science that we should not expect there to be a plethora of complete theories and explanations. Progress is all we have a right to demand, and surely a characterisation of processing that generates information that intuitively appears to be relevant to, or useful in, the formation of beliefs by means of the exercise of a cognitive capacity counts as progress.

<sup>21</sup> Of course Fodor's version of RTM does make it intelligible how subpersonal representation-manipulating processes could play a role in the fixation of belief or "make their mark" at the personal level. However, Fodor's theory is not the only model currently on the market. Consider a famous passage from Dennett.

In a recent conversation with the designer of a chess-playing program I heard the following criticism of a rival program: "It thinks it should get its queen out early." This ascribes a propositional attitude to the program in a very useful and predictive way, for as the designer went on to say, one can usually count on chasing that queen around the board. But for all the many levels of explicit representation found in that program, nowhere is anything roughly synonymous with "I should get my queen out early" explicitly tokened. The level of analysis to which the designer's remark belongs describes features of the program that are, in an entirely innocent way, emergent properties of the computational properties that have "engineering reality". I see no reason to believe that the

## 2.4 The nature of computation

In this section I will attempt to develop an account of the nature of computation. My account is largely in line with what might be described as the received view of computation, a widely held view that has been championed by Fodor.<sup>22</sup>

First and foremost computers are symbol manipulators; given symbols as input they produce symbols as output. Thus computational processes involve the manipulation or transformation of symbols. This symbol manipulation isn't random but rather rule governed. More specifically, computers generate output symbols from input symbols by applying symbol-manipulating rules (what this means will be explained in due course).<sup>23</sup>

Computational processes are often complex; that is to say that it is frequently the case that computers generate output from input not in one single symbol manipulating move but rather by executing a whole series of substeps each of which involve applying a rule to a symbol. However, to express the point in these terms fails to

---

relation between belief-talk and psychological process-talk will be any more direct. (1978, p. 107).

Even shorn of the implicit instrumentalism, such a view goes some way towards indicating how the manipulation of internal representations at the subpersonal level could play a role in belief fixation without beliefs being relations to internal representations.

Another alternative to Fodor's RTM identifies beliefs with whole collections of internal representational states. Horgan and Woodward (1985) portray Minsky's (1981) "Society of Mind" view in such terms. According to Minsky, they write, "the role of a belief (say) is typically played by a vast, highly gerrymandered, conglomeration of C[ognitive] S[cience]-events' (p. 140). In short then, scientific psychologists are not forced to endorse Fodor's RTM and there is no reason to suppose that they generally do.

<sup>22</sup> See 'Methodological Solipsism Considered as a Research Strategy in Cognitive Psychology' where Fodor makes his most explicit statement of his understanding of the nature of computation.

<sup>23</sup> Therefore, pace Cummins (1989), making hollandaise sauce is not computation.

distinguish between two distinct ways in which a computational process can be complex. The first type of complexity is the analogue of the complexity of the task of baking a cake when executed by a single individual.<sup>24</sup> The process of baking a cake generates a baked cake as output from a collection of ingredients as input. This process cannot be executed by an individual in one single move; rather she has to make a whole series of simpler moves in a certain order. Examples of such substeps would be those of breaking an egg into a bowl, mixing together some eggs with a quantity of flour, butter, and sugar, pouring a mixture of eggs, sugar, butter, and flour into a lined cake tin, and so on. A cook bakes a cake by performing each of these substeps in a certain order. The point to note is that the system that performs the complex task, namely the cook, is the very system that performs the simpler sub-tasks. A computational process that is complex in this first respect is one that is executed by a system by means of executing a whole series of simpler computational processes or operations in a certain order. In such a case the system that performs the substeps is the very system that executes the complex process and it does the latter by doing the former.

The second type of complexity is the analogue of the process of cooking meals as performed by a restaurant kitchen. A restaurant kitchen produces from an input of ingredients and representations of customer orders an output of meals which correspond to what the customers have ordered. In executing this process a whole series of simpler processes or operations are executed, but there is a respect in which it is misleading to say that the kitchen as a whole performs these simpler operations. Rather, it is subsystems of the kitchen that perform them. Restaurant kitchens have a hierarchical structure; that is, they can be analysed into subsystems each of which performs a distinct task (or series of tasks) the performance of which contributes to the higher level performance of the complete kitchen. These subsystems doing what they do, and interacting with one another in the process, engender the complex behaviour of the whole system to which they belong. In other words, the kitchen

---

<sup>24</sup> Examples from the world of food production are frequently used by philosophers and psychologists when explicating the notion of computation. See, for example, Cummins (1983) and Haugeland (1985). My love of food-preparation and consumption stands in the way of the generation of more original examples.



executes its complex process by having components that execute simpler processes and interact with one another in so doing. These subsystems in turn analyse into simpler subsystems - that execute simpler tasks and interact with one another in so doing - that analyse into simpler subsystems until eventually one comes to the basic components of the system. These basic components execute tasks that are complex, if complex at all, only in the first respect described above. The basic components of restaurant kitchens are typically individual humans that do such things as chop vegetables, stir sauces, carve joints of meat, arrange items of cooked food on plates, and so on. At any level in this hierarchy of systems there are operations simple relative to that level, simple in the respect that if they are performed by performing simpler operations those operations are performed by subsystems lower down in the hierarchy.

A computational process is complex in the second respect when it is executed by a computer with a hierarchical structure, so that it executes the process in question by having components that execute simpler tasks, and in so doing interacts with other components in that level of the hierarchy. In the case of a computer, the processes at all levels will be symbol-manipulating ones.<sup>25</sup>

Sometimes computational processes that are complex in the first respect occur in a computer with a hierarchical structure. This will be the case if at any level in the hierarchy there exists a subsystem that performs its task by executing symbol-manipulating substeps.

The concept of a program is intimately bound up with that of computation. What is a computer program? A program is a description of the means by which a computer executes a particular computational process. A program typically takes the form of a list of lines or a flowchart each line or box of which corresponds to a sub-step taken in executing the complex process in question, and thus to a symbol-manipulating capacity. The program as a whole will not

---

<sup>25</sup> The hierarchical nature of (many) computers is described by Fodor (1968), Dennett (1978), Lycan (1982) and Cummins (1983). All these writers emphasise this characteristic of computers in order to make a point about the explanation of computer capacities, the point being that explaining such capacities involves analysing the system under study into subsystems with simpler capacities the execution of which engender the execution of the target capacity.



merely specify the steps taken but also the order in which they are made. Each of the substeps will of course be a computational operation, that is, a symbol-manipulating move made by applying a rule to a symbol to generate from it another symbol. Thus each line (or box) of a program will correspond to a symbol-manipulating rule applied by the computer (or one of its subsystems) in executing a complex computational process. A program will rarely tell the whole story as to how a computer performs a given complex process, for typically each line of the program will correspond to an operation that is not absolutely simple but rather complex in at least one of the respects described above.

Programs are thus rather like recipes in cookbooks, for such recipes specify the operations to be made or the steps to be taken (and the order of execution) in executing the process of producing a certain dish. Each line of the recipe will specify a culinary operation which is not absolutely simple but simple only relative to the language of that cookery book; executing that operation may well involve executing a whole series of simpler operations in a certain order. Similarly, computer programs are typically written in a programming language (for example LISP or FORTRAN) and each line of a program will specify a computational operation simple only relative to that language. Thus in executing an operation specified by a single line of a LISP program a computer (or one of its subsystems) might have to execute a whole series of simpler computational operations.

This way of viewing computer programs<sup>26</sup> does not require a program to be explicitly represented in a computer that runs it. To run a program just is to execute a computational process by executing the steps specified by the program in that order. To do this no more requires the program to be explicitly represented in the computer than baking a cake in the manner described by a recipe requires the cook to consult some explicit representation of the recipe. In other words programs describe how computers perform computational processes, rather than specify the representations that are causally implicated whenever the process is executed. To say this is not to deny that programs are ever explicitly represented in computers. On the contrary, they often are, for programming a von Neuman machine typically involves bringing about changes in its memory so

---

<sup>26</sup> A way which is heavily influenced by Cummins (1983 ch. 2).

that it comes to contain representations that correspond to the lines of the program, representations that will be causally implicated in the machine's symbol-manipulating activity whenever it runs the program or executes the process that the program is a program of. But even computers that explicitly represent some of the programs they run cannot explicitly represent all of the programs that they run on pain of infinite regress. This is because a computer can only run an explicitly represented program if it "knows-how" to respond to those representations. If how it is to respond to such explicit representations has to be explicitly represented, then in turn how it is to respond to that explicitly represented program will need to be explicitly represented, and so on ad infinitum. To bring an end to the regress what is needed is what Dennett (1983) calls tacit representation. Tacit representation is 'know-how' that is 'built into the system in some fashion that does not require it to be (explicitly) represented in the system' (p. 218). If a computer didn't have such know-how then it wouldn't be able to respond to any explicitly represented programs. For a computer to have such know-how is for it to have certain computational capacities hard wired into it.

Computers are sensitive only to the formal properties of the symbols that they manipulate, being blind to such semantic properties as meaning, content, truth value, reference, and the like. Fodor expresses the point in the following terms:

computational processes are both *symbolic* and *formal*. They are symbolic because they are defined over representations, and they are formal because they apply to representations in virtue of (roughly) the *syntax* of the representations. . . . What makes syntactic operations a species of formal operations is that being syntactic is a way of *not* being semantic. (1980, 227)

Thus the rules that computers apply to representations are formal rules. Two computers, or computational processes, will belong to the same computational type if and only if they produce the same formally individuated symbolic output from the same formally individuated symbolic input, and do it by applying the same formal rules. Two computers or computational processes might meet this requirement despite the fact that the symbols that they manipulate

are divergent in their semantic properties. But this would not stand in the way of their computational equivalence. Alternatively, two computers, or computational processes, could produce semantically identical yet formally divergent output from semantically identical yet formally divergent input. In such a case, despite the semantic equivalence, the systems or processes would belong to different computational types.

How are we to understand this term "formal"? Syntactic properties are a species of formal property, but not all formal properties are syntactic. In this context the term "formal" can be replaced with "syntactic", a substitution that aids clarity given the familiarity of the latter notion. This substitution can be made because most scientific psychologists take the brain to be a type of computer that (like a von Neuman machine) is sensitive to syntactic properties, and they therefore understand the computational operations that underlie cognition to be syntactic operations. Thus for our purposes, computation can be understood as the process of manipulating syntactically individuated symbols by means of the application of syntactic rules.

What does it mean to say that computers apply syntactic rules to symbols? A syntactic rule, or a representation of a syntactic rule, defines or specifies a function the arguments and values of which are syntactically individuated symbols. Suppose a syntactic rule  $R$  defines a function  $F$ . A computational system or process  $C$  generates its output from its input by applying  $R$  to its input if and only if its input-output behaviour satisfies  $F$ : in other words, if and only if it takes arguments of  $F$  as input and whenever it takes such an argument as input, it produces as output the value of  $F$  for that argument. This characterisation of what it is to apply a syntactic rule indicates that a system or process can manipulate symbols by applying such rules without having either to represent them explicitly or to understand them.

It is because of the fact that computers manipulate syntactically individuated symbols by applying syntactic rules to them that they are described as syntactic engines.<sup>27</sup> But what about this term "syntax"? What are syntactic properties? For a symbol to have syntactic properties it must belong to a language. In this connection

---

<sup>27</sup> The term "syntactic engine" is due to Dennett (1981).

the term "language" is best thought of as referring to formal languages rather than natural languages. Formal languages are defined without reference to meaning or any other semantic property. A formal language consists of a set of sentences of well formed formulae (wff's for short), typically denumerably many such sentences. Defining a formal language involves specifying the basic or atomic symbols of the language (the words of the language), making clear to which syntactic category each such symbol belongs (the categories being: predicate symbol, constant, quantifier, logical connective, and so on) and a set of formation rules that determine which combination of words of the language are wff's. Symbols can be combined in accord with the rules to create complex expressions of the language. Such sentences can similarly be combined in accord with the rules to create further, more complex, sentences. Given that one can continue combining sentences in this manner to create new ones, a formal language with only a small number of words and formation rules will comprise denumerably many distinct sentences or wffs.

Describing the syntactic structure or the syntactic properties of a complex wff (that is, one made from combining simpler wff's) is a matter of specifying the simpler wff's that comprise it and the manner in which they are combined. Similarly, describing the syntactic structure of a simple wff (that is, one that cannot be analysed into simpler wff's) involves specifying the words that are its constituents and the manner in which they are combined. Thus specifying the syntactic properties of a wff involves analysing it into simpler components and describing how those components are combined. Ultimately these components will be words of the language.

Thus the notion of syntax is inextricably bound up with that of a language; you don't have syntactic properties where you don't have language. Moreover, syntactic properties are inextricably bound up with semantic properties despite the frequently highlighted contrast between them. This is so for several reasons. Firstly, it is part of the essence of symbols that they are meaningful items, so something has syntactic properties only if it has semantic properties. Secondly, a property of a symbol is a syntactic property only if it has a bearing on the symbol's meaning. As Dennett puts it, 'what makes a feature



syntactic is its capacity to make a semantic difference' (1982, p. 141). Thus the colour or size of a printed word is not a syntactic feature of it. Thirdly, to assign a symbol to a syntactic category is to imply something about its meaning. For example, to say that a particular symbol is a two-placed predicate-symbol is to say that it expresses a relational property that, when instantiated, holds between two individuals. (Bermudez, 1995).

The close relationship between semantic and syntactic properties might appear to put pressure on the claim that computers are sensitive only to syntactic properties by suggesting an argument such as the following. For any computer there will be a systematic relationship between the semantic properties of the symbols that it manipulates and the syntactic properties of those symbols; for example, semantic differences between symbols will be reflected in syntactic differences between them. Consequently, corresponding to any counterfactual-supporting syntactic generalisation true of a computer there will be a counterfactual supporting semantic generalisation. What this entails is that how a computer processes its input will depend upon the semantic properties of that input in the respect that: (i) *ceteris paribus*, to have been processed differently the input would have to have been semantically different (for only then would it have been syntactically different); and (ii) *ceteris paribus*, had the input been semantically different it would have been processed differently (as it would have been syntactically different). All this makes it difficult to avoid the conclusion that if computers are sensitive to syntactic properties then they are just as sensitive to semantic properties.<sup>28</sup>

One reply to this argument is that it supplies the ammunition for its own downfall by implying that computers are sensitive to semantic properties in virtue of their sensitivity to syntactic properties (and not vice versa). This dependency relationship suggests that, first and foremost, computers are syntactic engines. However, replying in this manner is to run the risk of winning the battle at the cost of losing the war. Many computers (and the symbols that they manipulate) are physically realised. The behaviour of such systems is governed by the

---

<sup>28</sup> A similar argument is run by Block (1989) in an attempt to defeat the claim that the semantic properties of intentional states are epiphenomenal with respect to the intentional states and behaviour that they cause.



laws of physics. Hence any sensitivity to the syntactic properties of the symbols that they manipulate will be a product of their sensitivity to the physical. In other words, the dependency relationship between the syntactic and the physical will mirror the one that holds between the semantic and the syntactic. So, by parity of reasoning, many computers - including those that psychologists care about - are physical rather than syntactic engines.

However, the claim that computers are syntactic engines, that they are only, or at least primarily, sensitive to the syntactic properties of the symbols that they manipulate, can be defended.

Consider the case of neurophysiological systems. Some subsystems of the human brain take symbolic input and produce symbolic output, the symbols so manipulated having semantic properties. There will be semantic generalisations true of the input-output behaviour of such neurophysiological systems, generalisations of the form: if the system receives input with such and such semantic properties then it will respond by producing output with such and such semantic properties. If this is not the case then scientific psychology is a non-starter. Yet it offends intuition to claim that such neurophysiological systems are sensitive to semantic properties; intuitively it seems plausible to say that neurophysiological systems are sensitive only to neurophysiological properties.

The question is: how can we reconcile the view that there are semantic generalisations true of (some) neurophysiological systems with the view that such systems are sensitive only to neurophysiological properties? Here is how. Neurophysiology individuates events, states, processes and entities in terms of their neurophysiological properties. Semantic properties cannot be identified with, or reduced to, such properties. A taxonomy of neural phenomena in terms of their semantic properties would crosscut a taxonomy of neural phenomena in terms of their neural properties. This is because neural phenomena belonging to one and the same neural type can diverge in their semantic properties and, conversely, neural phenomena with the same semantic properties can differ at the neural level. Consider a neurophysiological system *N* located in my brain. *N* will take input and produce output that has both semantic and neurophysiological properties. Moreover, there will be true of *N* both semantic and neurophysiological generalisations

concerning its input-output behaviour. Now just the same neurophysiological generalisations will be true of the input-output behaviour of all neurophysiological systems of the same type as N. But there is no guarantee that the same semantic generalisations will be true of them all. Thus there may be a respect in which N is sensitive to the semantic properties of its input. But given that there will not be any (or need not be any) semantic generalisations that apply to all neurophysiological systems of the same type as N we can conclude that, qua neurophysiological system, N (and all other neurophysiological systems) is not sensitive to semantic or syntactic properties.

A parallel argument establishes the conclusion that computers are not sensitive to semantic properties. Computational phenomena are individuated in terms of syntactic properties. Semantic properties neither reduce to, nor can be identified with, syntactic properties. It is possible for the symbols manipulated by one computer to have quite different semantic properties than the corresponding symbols manipulated by another computer despite the fact that the two computers are computationally identical. Thus all systems of a particular computational type will have the same syntactic generalisations true of them but not the same semantic generalisations. Therefore, we can conclude that computational systems, qua computers, are not sensitive to semantic properties but rather are sensitive to syntactic properties.<sup>29</sup>

That completes my account of the nature of computation. By way of recapitulation: computers are symbol manipulators, they produce symbolic output from symbolic input. The symbols that they so manipulate have syntactic properties (and thus belong to a formal language) and they are manipulated by means of the application of

---

<sup>29</sup> This argument does not presuppose an externalist theory of content in endorsing the possibility that syntactically identical symbols manipulated by computationally identical systems can diverge in content. This is because computational systems can be embedded in larger computational systems. Consider two computationally identical systems S' and S'' embedded in computationally divergent systems. The causal relations between the symbols manipulated by S' might diverge from those between the same symbols manipulated by S'' (in virtue of the differences between the respective systems in which S' and S'' are housed) in such a way that, on a functionalist theory of content, their meanings diverged.

syntactic rules. Despite the fact that symbols have semantic properties (and often physical properties) computers are sensitive only to the syntactic properties of the symbols they manipulate.

## 2.5 How could a physical system be a computer?

Some physical systems are computers. According to scientific psychology the brain, or some of its parts, is a computer. Thus the question arises of how a physical system could be a computer or engage in computational activity. What conditions must a system satisfy in order to be a computer? I will attempt to answer this question in this section.

Suppose we are presented with a physical system (call it S) which is described as a computer. For this description to be legitimate - given the above account of the nature of computation - there must be some language that S employs, a language the symbols of which S manipulates. Suppose that a description of that language (call it L) is given by specifying its atomic symbols, or words, and the formation rules for constructing sentences, or wff's, out of these words. For S to employ L there would have to be a mapping of potential internal physical states of S onto symbols of L, a mapping that satisfies two conditions. The first condition is this. The mapping maps onto each word and expression of L a distinct type of internal state that S is capable of tokening. The physical states mapped onto the complex symbols must be more complex than the physical states mapped onto the atomic symbols and the complexity of these complex physical states must mirror the complexity of the symbols onto which they are mapped. Let me try to explain. Complex expressions of a language are made up of simpler items, ultimately words, that are combined in a certain way. Tokens of such complex symbols have components each of which is a token of a word type present in the complex symbol-type in question. For example, when I say "Giraffes have long necks" what I do is emit a sound that is a token of the sentence type *Giraffes have long necks*. This sound can be decomposed into parts each such part being a token of a word that belongs to the sentence. If, for example, the noise I made did not have a component which was a token of the word *Giraffes* then whatever sentence I uttered it was not *Giraffes have long necks*.

In the case of the mapping of physical states of S onto symbols of L, the physical states that are mapped onto complex symbols must have as components physical states that map onto the simple symbols that are the constituents of the complex symbols. Thus, for example, if L is English, then the physical state of S mapped onto the sentence *Giraffes have long necks* must be a complex state which has as components the physical state that is mapped onto *Giraffes*, the physical state that is mapped onto *long* and so on. This condition would be violated if, for example, the state mapped onto *Giraffes have long necks* had no component that mapped onto *Giraffes* but had one which mapped onto *tree*.

What makes a sentence the sentence that it is is not just the words that it comprises but, in addition, the way in which those words are combined to make the sentence, in other words the syntactic structure of the sentence. This fact generates a second condition that the mapping must satisfy if S is to employ L. The physical relationships between the components of complex states must mirror the syntactic relationships between the words of the sentences onto which they are mapped. Consider an example. The state mapped onto *Aardvarks eat termites* must have the same simple states as components as that mapped onto *Termites eat aardvarks*, but the physical relationships between these simple components must differ. They must differ in the respect that the component of the first state that corresponds to *aardvark* had better not stand in the same relationship to the other components of that state as the component corresponding to *aardvark* in the second state stands to the other components of that state. However, there is a respect in which the physical relationships between the parts of these physical states must agree. Syntactic relationships must correspond to physical relationships so that all sentences that have the same abstract syntactic structure will be mapped onto complex states whose components stand in the same physical relationships to one another. For example, all sentences of the form A eats B must have mapped onto them complex states such that the component that corresponds to A stands in just the same physical relation to the components that correspond to the other words in the sentence. In meeting this second condition, a mapping of physical states of S onto symbols of L effectively maps physical relations between component physical



states of complex physical states onto syntactic relationships between symbolic components of complex symbols.

For S to employ L there must be a mapping from internal physical states of S onto symbols of L that satisfies the above described conditions. But how is a theorist supposed to specify a proposed mapping? Given that languages standardly comprise denumerably many distinct sentences, one clearly cannot just write down for each and every distinct sentence of L the state that maps onto it, for to attempt to do this would be to engage in a task that could never be completed. But there is a device that can be employed that is analogous to that employed in defining a formal language. Formal languages are defined by specifying the atomic symbols - the words - of the language, and the formation rules which determine which combinations of words are sentences of the language. Similarly, in specifying a mapping from physical states to symbols one explicitly states for each word of the language the physical state that maps onto it, and for each distinct syntactic relation that can hold between components of a sentence the corresponding physical relation. Once this has been done the theorist has effected a mapping of physical states of S (that is, states S can potentially token) onto symbolic expressions of L that satisfies the above described conditions.

A first step in justifying the assertion that a particular physical system is a computer is to indicate which language it employs and specify a mapping of internal physical states of the system onto symbolic expressions of the language in question. Once this is done one has indicated how the symbolic expressions of the language are realised or instantiated in the system in question. Given this fact I will, following Pylyshyn (1984), name such a mapping of physical states onto symbolic expressions an "instantiation function".

It is possible for physically distinct systems to employ the same language. In such a case the instantiation function that describes how the symbols of the language are realised in one system will differ from that that describes how the symbols of the language are realised in the second system. Thus the symbolic expressions of a language are multiply realisable in the respect that if a physical system employs a particular language then there will be some other (possible) physical system that employs the same language yet realises the symbols of that language differently at the physical level. Indeed, the



instantiation function true of a system can change over time with symbols being realised in one way at  $t_1$  and quite another at  $t_2$ .

I do not wish to argue that if there is an instantiation function that maps the states of a physical system  $S$  onto the symbols of a language  $L$  then  $S$  will thereby be a computer, or more specifically, a computer that employs  $L$  or manipulates the symbols of  $L$ . Satisfying this condition is not enough to attain computerhood. In other words the existence of such a mapping is not a sufficient condition for computerhood. But it certainly is a necessary condition. So what other conditions are there?

A first further condition has to do with meaning. Earlier we saw that there is an intimate connection between semantic and syntactic properties. Items that do not have semantic properties are not symbols, and a property of an item is a syntactic property only if it has a bearing on the meaning of the item. Consequently, for a physical system to be a computer, some of its internal states (namely, those mapped onto the symbolic expressions of a formal language by whatever instantiation function the system satisfies) must have semantic properties.<sup>30</sup> One could imagine there being a physical system such that it was possible to map its internal states onto the symbols of a formal language despite the fact that those internal states had no semantic properties. Such a system would not be a computer.

However, it is not enough that the putative symbols manipulated by a physical system have semantic properties. In addition, there must be a systematic relationship between their semantic and syntactic properties; a relationship like that that holds between the semantic and syntactic properties of natural language symbols. A relationship such that, for example, the meaning of any complex symbol is determined by its syntactic structure and the meaning of its parts.

Another condition is that to be a computer a system's putative symbol-manipulating activity must be semantically coherent. This condition would not be satisfied if the system generally produced output which bore no sensible or cogent logical or semantic relation

---

<sup>30</sup> These semantic properties could be the products of acts of interpretation of intelligent agents or, alternatively, the products of the satisfaction of some naturalistic sufficient condition for the tokening of such properties.

to its input. In such a case the system would, as it were, "talk gibberish" and would not be doing anything that would count as information processing, solving information processing problems, or extracting information.

In making these points I am not turning my back on the syntactic account of computation that I advanced earlier. For though computation, by its very nature, involves the manipulation of meaningful items and the generation of output that is semantically related to the input from which it is generated, computers and computational processes are individuated syntactically rather than semantically.

I feel tempted to claim that the notion of a computer is partly a teleological one; that nothing is a computer unless it is its function to process or extract information, solve problems, work things out, and such like. Computers don't just do these things; in addition they are used to do these things and in so being used are of a benefit to their user. With respect to the brain this suggests the following. A subsystem of the brain might satisfy all of the above described conditions but if it doesn't benefit the system in which it is housed in virtue of generating the information that it generates (if it isn't, so to speak, used to generate that information or if that information doesn't play a significant role in the life of the embedding system) then it isn't a computer.

This attempt to specify the conditions that a physical system must satisfy in order to be a computer could do with much in the way of elaboration. However, I think that I have done enough to indicate the outlines of an adequate answer to the question of how a physical system could be a computer, an answer which tells us in virtue of what those physical systems that are computers attain that status. What should be clear is that it is very hard to be a computer, contrary to what some philosophers would have us believe.

Suppose that physical system *S* is a computer that manipulates symbols of *L*. What computational capacities will *S* have, what symbolic/syntactic functions will it be able to compute? The answer is that it all depends upon what counterfactual-supporting generalisations concerning its internal state transitions are true of *S* (or its parts). Given the instantiation function, corresponding to each such generalisation will be a syntactic generalisation. Thus if it is true

of S that whenever it tokens an internal state of type I' it subsequently tokens an internal state of type I'', then it will also be the case that whenever it tokens the symbol F' it will generate from it a token of the symbol F'' (where F' and F'' are the symbols that have, respectively, I' and I'' mapped onto them). Given the huge network of generalisations relating its internal physical states to one another, there will be a huge network of syntactic or symbol-manipulating generalisations true of S. Which such generalisations are true of S will determine which symbol-manipulating capacities it has, or which symbolic functions it is capable of computing. Suppose that it is claimed of S that it can compute the symbolic function SF, a function defined by the rule R. This claim will be true if and only if the syntactic generalisations that are true of S are such that whenever S tokens a symbol that is an argument of SF, that token causes S to token the symbol that is the value of SF for that argument (or, alternatively, S responds to the token by producing a token of the symbol that is the value of SF for that argument). If this condition is satisfied, some of S's (potential) symbol manipulating activity can be described as generating symbols of L from symbols of L by applying rule R.

We have seen that syntactic properties or syntactic types are multiply realisable. But is it true to say, as does Block (1989), that 'syntax is . . . a functional notion', and that 'it is having a certain functional role that makes a state satisfy a *syntactic* description' (p. 142)? For syntactic properties to be functional properties it would have to be the case that a state of a system's having a given syntactic property was a matter of its bearing certain counterfactual-supporting causal relations to other states of the system. A consideration of symbols containing logical connectives would seem to support Block's view. Most powerful formal languages contain logical connectives, and such connectives are a primary way by means of which simple sentences are combined to build more complex sentences. Consider the connective &. For a state of a system to have the syntactic property of being a sentence of the form  $A \ \& \ B$  certain causal generalisations would have to be true of the system. Suppose that the state in question is state I' of system S. For I' to have the syntactic property of being a symbol of the form  $A \ \& \ B$  it would have to be the case that the following causal generalisations held. First, I'

causes the state that maps onto  $A$ , (or realises  $A$  in  $S$ ), and the state that maps onto  $B$  (or realises  $B$  in  $S$ ). In other words, whenever  $S$  tokens  $I'$ , that token causes a token of the state that maps onto  $A$ , and a token of the state that maps onto  $B$ . Second,  $I'$  is jointly caused by the state that maps onto  $A$  and the state that maps onto  $B$ . In other words, whenever  $S$  tokens the state that maps onto  $A$  and tokens the state that maps onto  $B$ , those two tokens jointly cause a tokening of  $I'$ .<sup>31</sup> These generalisations correspond, respectively, to the familiar rules of  $\&$ -elimination and  $\&$ -introduction. The same will hold, *mutatis mutandis*, for all other connectives; that is, for any connective  $C$ , for a state of a system to have the syntactic property of being a symbol containing  $C$  as its main connective, generalisations corresponding to the introduction and elimination rules for  $C$  must be true of the system.

Indeed something like the above would also appear to apply to quantifiers. For a state of a system to have the syntactic property of containing a particular quantifier  $Q$ , that state must figure in causal generalisations that correspond to the introduction and elimination rules for  $Q$ .

However, we have to be careful, for it is certainly possible for two distinct computational systems, or the same system at different points in time, to employ the same language, yet to manipulate the symbols of that language differently. In such a case the causal relations between the symbolic states of one of the systems will diverge from the causal relations between the symbolic states of the other. Many of the syntactic generalisations that are true of a computational system are contingent in the sense that those generalisations don't have to be true of the system given the formal language that it employs; it is consistent with a computer's employing a particular language that it manipulates the symbols of that language in many different ways. Indeed, what happens when one programs a computer is that one brings about changes in the way it manipulates the symbols of the language that it employs; in other

---

<sup>31</sup> Strictly speaking this is a little too strong as  $S$  need only tend, or have a tendency to, generate a token of the state that maps onto  $A$  and a token of the state that maps onto  $B$  from tokens of  $I'$ , and a tendency to generate tokens of  $I'$  from tokens of the state that maps onto  $A$  and the state that maps onto  $B$ .



words one alters the syntactic generalisations that are true of the computer.

## **2.6 The role of meaning in psychological explanation**

As we have seen, computation and meaning are closely linked in virtue of the fact that computation involves symbol manipulation. The concept of symbol and that of meaning are inextricably bound together. It's not just that symbols are sometimes, or often, meaningful. Rather, to be a symbol you have to have meaning, or belong to a system of meaningful items (that is, a language), or have been designed or invented with a view to having meaning attributed, or something along those lines. Thus no computation without meaning. However, this fact alone does not entail that psychological explanation must appeal to meaning or semantic properties. That computation, by its very nature, involves the manipulation of meaningful items doesn't entail that computational processes are to be individuated in terms of the meanings of the symbols they manipulate, or that one's description of a computational process qua computational process is incomplete until one has specified some meanings. Indeed, I have argued that computational phenomena are syntactically individuated. Thus one can describe a computer's computational capacities and activities qua computational without concerning oneself with the meaning of the symbols the system manipulates, even if those symbols must have some meaning to be symbols.

This might suggest the following line of thought. Scientific psychology is committed to the idea that the mind-brain - or its subsystems - is a computer, and thus that the processes and capacities that underlie cognition are computational. In accounting for a cognitive capacity, psychologists attempt to specify the program that is run whenever that capacity is exercised. Doing this involves describing a series of symbol-manipulating steps each of which corresponds to a symbol-manipulating capacity of the system under study (or one of its subsystems). As meaning has no place in the description of a computer's computational operations and capacities, meaning, and semantic properties in general, have no place in the explanation of cognitive capacities.



However, I shall argue, meaning has a fundamental role to play in psychological explanation. A description of a computer program or a series of computational capacities that made no attributions of semantic properties to the representations it alluded to could not constitute a satisfactory explanation of a cognitive capacity.

The nature of computation would be such as to imply that there is no role for meaning in psychological explanation only if it were the case that cognitive capacities were computational or syntactic capacities. But the crucial fact is that scientific psychology is concerned with explaining intentionally-characterised cognitive capacities. Scientific psychology characterises and individuates cognitive capacities (partly) in terms of the intentional states that their exercise gives rise to. A capacity the exercise of which did not standardly result in the formation of a belief about the nature of the world by means of vision would not be the capacity Marr was interested in; a capacity the exercise of which did not standardly result in the formation of a belief about the meaning of a sentence just heard would not be the capacity to understand spoken sentences; a capacity the exercise of which did not standardly result in the having of a recollection of a past event would not be the capacity to remember past events; and so on. A consequence of this fact is that even if computation is involved in cognition, explaining or describing a system's computational capacities would not be enough to explain its cognitive capacities. At the very least the meaning or semantic properties of the symbols manipulated would have to be specified.

It is a fundamental assumption of scientific psychology that the symbols manipulated by the subpersonal systems of the mind-brain have semantic as well as syntactic properties and that the semantic properties of these symbols are crucial to our having the cognitive capacities we have. The idea is that it is by means of computation that the mind-brain processes information or solves information processing problems. Thus, for a computational process or capacity to be involved in the exercise of a particular cognitive capacity, the symbols manipulated would have to have appropriate meanings; not any old meaning will do. Hence, alongside the instantiation function which will map neurophysiological states onto syntactically-individuated symbols of a formal language, there will be a semantic

function which maps such syntactically-individuated symbols onto meanings. This semantic<sup>32</sup> function will be such that the meaning of complex symbols will depend upon the meaning of their parts and their syntactic structure. In other words, the semantic function will reveal the language(s) manipulated by the mind-brain (and/or its sub-systems) to have a combinatorial semantics. It is important to stress that, as far as the scientific psychologist is concerned, the semantic function maps mental symbols onto meanings that they really have. Thus the content of mental symbols is not "as if" content or the product of interpretation. If the semantic properties of mental symbols and states were not as real and objective as, say, their physical properties, then whatever goes on within the mind-brain at the subpersonal level would not be information processing.<sup>33</sup>

Given the existence of a semantic function, whenever a computational capacity is exercised by a sub-personal system, a symbol with a certain meaning will be generated from a symbol with some other meaning. The generation of such a symbol will constitute an important step in a process that will eventuate in the formation of a personal level intentional state that manifests some cognitive capacity or other. But how and why will making such steps, steps which involve the generation of a symbol with a certain meaning from a symbol with some other meaning by means of computation, play such an important role in the exercise of cognitive capacities? To answer this question consider the example of my capacity to work out how much money I have in my bank account.

Most of the time I don't know how much money I have in my bank account but I have the capacity to work it out. How do I do this? First of all I consult my chequebook which has in it a symbol that represents the state of my bank balance at some previous point in time and a whole load of other symbols each of which represent individual deposits and withdrawals from my account in the

---

<sup>32</sup> In naming this function "the semantic function" I am following Pylyshyn's (1984) terminology. Block (1990) uses the same term to label functions from meanings to meanings.

<sup>33</sup> Therefore, McDowell's (1994) account of scientific psychology as being a syntactic psychology whose attributions of content to subpersonal states are not literal is not one that the practitioners of scientific psychology would recognise as an accurate description of their discipline.

intervening period, and which collectively represent all such deposits and withdrawals. These symbols provide me with a whole body of information from which I can work out how much money I have in my account now. Hence, what I do is generate information concerning the current state of my bank account from information about certain events in its history. I make this transition by working out, or extracting, information that I want from information that I already have by means of computation. In other words, I execute a whole series of symbol manipulating moves, and in so doing generate the information that I require from information that I already have. These symbol-manipulating moves are the following:

- (i) On symbols which represent all the deposits I have made I perform a symbol-manipulating operation which generates a symbol that represents their sum.
- (ii) On symbols that represent all the withdrawals I have made I perform a symbol-manipulating operation that generates a symbol that represents their sum.
- (iii) On the output symbol of (i) and that of (ii) I perform an operation that generates a symbol that represents the net deposit.
- (iv) Finally, on the symbol that represents the previous state of my bank account and the output of (iii) I perform an operation that generates a symbol that represents the current state of my bank balance.

In executing each of these steps I perform, quite mechanically, a computational operation which generates symbolic output from symbolic input. The symbols that I manipulate belong to the Arabic numerical system and so in addition to their syntactic properties they have semantic properties, in particular they represent numbers. In virtue of their representing numbers in performing the computational operations that I perform, I compute the value of a mathematical function for given arguments in performing each of (i) to (iv). At stages (i), (ii) and (iv) the mathematical function is that of addition whereas at (iii) it is that of subtraction. But, as indicated in the description of the stages, the symbols don't just represent numbers but also facts about my bank account or events in its history. The great thing is that given the meanings of the symbols that I

manipulate I can, by performing mechanical symbol-manipulating operations on them, work out the value of mathematical functions for given arguments and in so doing discover facts about my bank balance.<sup>34</sup> At each stage I discover a new fact about my bank account and its history, a fact from which, by making the appropriate moves - ultimately quite mechanical moves - I can work out either the fact I am after or a fact that it is helpful to be acquainted with if one is eventually to become acquainted with the target fact. For my symbol manipulating moves to generate information, it is not necessary that I be aware that they generate information or be aware of what information they generate. For example, the output of stage (i) represents the total value of deposits whether I realise it or not. One can be in the position of Searle in the Chinese room and still be generating information or uncovering facts.

In working out my current bank balance I manipulate symbols by applying syntactic rules to them. However an attempt to describe how I perform this task which only appealed to the syntactic properties of the symbols I manipulate and the syntactic operations that I apply to them will at best only describe how I compute a certain syntactic function. It would, to put it mildly, be incomplete as an account of how I work out my bank balance. It would leave it a complete mystery as to why performing those syntactic operations was relevant to the task at hand and why their execution reliably leads to success. For it is in virtue of their generating relevant information that these syntactic operations enable me to work out my bank balance.

A complete account must attribute meaning to the symbols manipulated, and specify the facts that the symbols at different stages of the process represent, for, *ceteris paribus*, had those symbols had

---

<sup>34</sup> It is quite common for a computer to compute a mathematical function and thus for its behaviour to be mathematically characterisable. To so characterise a computer's behaviour is to commit oneself to the claim that the symbols manipulated by the computer in question represent mathematical objects (for example, numbers). As in the bank account case, symbols that represent numbers will often do double representational duty representing the numerical magnitude of some feature of the non-mathematical world. The behaviour of two computers might, from the mathematical point of view, be identical despite the fact that they generate quite different information.



different meanings or represented different facts they would not have been relevant to, or helpful in, the execution of the task at hand. A crucial feature of the steps that I execute is that each of them constitutes the discovery of certain information or a certain fact, information or a fact that it is helpful to uncover as one can work out from it, by mechanical means, some other fact relevant to the task at hand. I work out the sum of the deposits and the sum of the withdrawals. Why do I do that? What is the point of doing that? If I discover such information, which I can do from the information that I started with, I can subsequently work out the net deposit. And from this, along with the information I started with, I can, by quite mechanical means, work out my current balance, which is my ultimate aim.

That I can acquire relevant information by applying syntactic operations to symbols that encode some other information depends to a large extent on the world's being a certain way, a way that it need not have been. For example, I can work out the value of the total withdrawals from my account by carrying out that symbol-manipulating operation by means of which I compute the sum of their individual values, because of a fact about the way the banking world works; namely, that the total value of withdrawals from one's account equals the sum of the values of the individual withdrawals. One could imagine a world, not so far removed from ours, where this fact did not hold; where, for example, banks subtracted from one's account an extra ten per cent of the sum of the individual withdrawals if your middle name began with the letter "J". Given the importance of these facts about the world, they must be appealed to in a complete account of how I work out my bank balance for if they are not specified it will be unclear as to why doing what I do is relevant to the task at hand and a means of generating useful information.

In short, then, an account of my capacity to work out my bank balance or, in other words, a description of how I do it, must appeal to more than the syntactic properties of the symbols I manipulate and the syntactic operations I apply to them in exercising this capacity. In addition it must specify the semantic properties of those symbols and describe relevant features of the world. But why, it might be asked, must syntax figure in this account? The answer is



that an account minus the syntax would be incomplete for such an account would leave it a mystery as to how I get from one stage to the next, for example, how I work out the value of my total withdrawals from knowledge of, or information concerning, the value of each of the individual withdrawals.

What is the relevance of all this to the issue of the nature of psychological explanations of cognitive capacities? The explanations that psychologists seek are a lot like the above explanation of my capacity to work out my bank balance. It is a fundamental assumption of scientific psychology that our having the intentionally-characterised cognitive capacities that we have depends to a large extent on our having subpersonal systems that are capable of working out or generating relevant information. Consider the visual capacity. I have the capacity to see, that is, the capacity to form true beliefs about the nature of my immediate environment (or, alternatively, to have experiences in which the world is presented as being a certain way, a way that it in fact is) by means of vision. Exercises of this capacity are processes that begin with light hitting the retina forming a retinal image, and end with the formation of a belief or the having of an experience. The question is: how am I able to get from a pair of retinal images to true beliefs about the nature of the world? Put this way, we have what looks like a miraculous achievement. The key insight of scientific psychology is that the pairs of retinal images that are the inputs to the visual process contain information from which it is possible, to work out or extract information about the nature of the external world just as the input to the process of working out my bank balance contains information from which it is possible to work out my current bank balance. This information about the external world is worked out or extracted by a subpersonal system, namely the visual module, and is precisely the kind of information that it is necessary to work out if the visual process is to eventuate in true beliefs about the external world.

The subpersonal visual module doesn't produce true beliefs about the nature of the world as output, but what it does do is perform a task that is fundamental to the generation of such beliefs, namely that of extracting relevant information, information that is such that were it not extracted, *ceteris paribus*, we wouldn't be able to form any true beliefs about the nature of our immediate environment by

means of vision. In general then, the scientific psychologist holds that our brain houses a battery of subpersonal systems that perform information-extracting tasks that must be performed if we are to have the cognitive capacities that we have. Thus an essential component of the task of accounting for a given cognitive capacity is to discover how the relevant subpersonal system generates useful information from whatever information it receives as input. In the case of vision this will involve describing how the visual system generates from retinal images - which contain information about the intensity of light falling on the retina - information concerning the nature of the subject's immediate environment. In the case of the capacity to understand spoken sentences it will involve describing how information concerning the semantic and syntactic properties of a heard sentence is generated from information concerning the nature of a sound that impinges on the auditory system.

Thus the subpersonal processing that underlies cognition is like working out one's bank balance in that it involves working out or generating information from information. Indeed there are further points of similarity. First, the information is symbolically encoded so that subpersonal systems generate information from information by generating symbols from symbols. Second, typically the output information will be generated from the input information in a series of moves each of which involves working out information from information that the system in question already has. This information will be relevant to the task at hand in that it will be the kind of information from which one can extract the information one desires or information that gets one nearer to that final goal. Third, these information-extracting moves are made by means of computation, that is, by applying syntactic rules to symbols. Thus subpersonal systems work out what they work out by applying syntactic rules to information-bearing symbols so as to generate further information-bearing symbols.

Given this similarity between the subpersonal activity that underlies cognition and the process of working out one's bank balance, semantic properties of symbols must be appealed to in describing the former for just the same reasons as they must be appealed to in describing the latter.

Suppose that in attempting to specify how a subpersonal system performs a given information-generating task a psychologist produces an account which appeals only to the syntactic properties of the symbols the system in question manipulates, and the syntactic rules that it applies to them. This account will be inadequate as it will leave it a complete mystery as to the relevance of all this computational activity. What is the point of making all those symbol manipulating moves? Why are they helpful in generating the information that the system generates? Why don't they support some other cognitive capacity? In order to answer these questions it is necessary to specify the semantic properties of the symbols manipulated, and thus to reveal the intermediary steps taken as being those of working out information that is relevant to the achievement of the overall goal of the system. There is only some value or point in executing a particular symbol-manipulating operation if by making it one can generate information which it is worthwhile for, or relevant given, the task at hand. To ignore the semantic properties of the symbols manipulated is thus to ignore what is most important about them and the operations that are applied to them.

Another good reason for the scientific psychologist not to ignore semantics is that it enables her to capture generalisations and recognise similarities that do not exist at the syntactic level. For example, from the syntactic point of view there are different ways of computing such arithmetical functions as addition and subtraction. One way involves the application of syntactic rules to Arabic numerals whereas another involves the application of quite different syntactic rules to binary strings. Consider two calculating devices, one that calculated in the first way and the other in the second way. From the syntactic point of view their respective capacities are completely different and their input-output behaviour is not subsumed by the same generalisations. It is only when the semantic properties of the symbols manipulated by these devices is considered that the salient similarities between them can be recognised; in particular, the fact that they both compute the same arithmetical functions. This gives rise to the worry that a purely syntactic psychology (of the sort championed by Stich (1983)) would employ a taxonomy that is far too fine-grained for its own good.

Syntactic capacities and operations do not appear from nowhere. When a system has a particular syntactic capacity or engages in a particular syntactic operation there is usually a good reason for its having that capacity or engaging in that operation. That good reason has to do with semantics. Consider the class of human artefacts that are computers. Their syntactic capacities and behaviour is a product of the way in which we have designed and programmed them. We have designed and programmed them in the way that we have precisely because computers so designed and programmed generate information - or have the capacity to generate information - that we particularly care about and want generated. In short, it is our desires and needs at the semantic level that dictate how we design and program computers at the syntactic level. A computer that didn't generate information that we needed or wanted would be of no interest or use to us and thus would not have come into being (save by accident).

Something similar is true of biological systems that are computers or have computational components. Syntactic capacities and processes get selected for in virtue of their generating information which is of use to, or beneficial for, the system that has the capacity or executes the process in question. Evolutionary forces tend not to produce creatures with syntactic capacities the exercise of which generally does not issue in significant semantic benefits. Consequently, a scientific psychologist who was interested only in the syntactic would be ignoring what was, from the evolutionary point of view, most important about the syntactic capacities and processes she sought to describe.<sup>35</sup> This point further emphasises the close relationship between syntax and semantics, and reinforces the claim that scientific psychology cannot fruitfully view the mind (or its parts) as a computer without taking an interest in the semantic properties of the symbols that it manipulates.

There are also powerful pragmatic reasons for the scientific psychologist to concern herself with the semantic properties of

---

<sup>35</sup> This suggests that teleological theories of content have got matters the wrong way round. Mental representations do not have the content that they have because of what they (or the processes that generate them) have been selected for; rather, it is because of the content that they have that they (or the processes that generate them) are selected for.



mental representations. Even if she set herself the task of uncovering the syntactic workings of, say, the visual module (rather than that of explaining an intentionally-characterised cognitive capacity) she could not afford to ignore the semantic properties of the symbols manipulated by that module. This is because one needs to have a pretty good idea of what information the visual module is processing before one has a realistic chance of working out any of the syntactic details. In general, in order to uncover the syntactic workings of a cognitive module psychologists proceed by first addressing such questions as: what information would it be useful for the module to generate for it to support the cognitive capacity that it underlies? What information could it possibly generate given the nature of the external environment? And so on. It is only when such questions have been answered (when some light has been shed on the semantic activity of the module in question) that the syntactic details can be uncovered. In the absence of a semantic story, the psychologist will not know what she is looking for and will have no clues as to how to formulate and test plausible hypotheses. Of course, it doesn't follow from this that semantic properties have a role to play in the description of the syntactic workings of a cognitive module or in an explanation of such a module's syntactic capacities. For such purposes the semantic properties hitherto identified should be, as it were, rubbed out. But they cannot be rubbed out from an explanation of an intentionally-characterised cognitive capacity for the reasons that I have outlined above.

In short, then, the semantic properties of the symbols manipulated by subpersonal systems cannot be ignored by the psychologist given her explanatory ambitions. But why must any appeal to syntax and syntactic operations be made? What would be wrong with an account which merely analyses a subpersonal system's information-generating activity into a series of information gathering substeps along these lines: the system starts off with such and such information from which it extracts some other information, from which it extracts some more information, and so on? The answer is that without the syntactic details it is a mystery as to how the system works out the information that it works out. As it generates information by means of computation, to give a complete account of how a system generates the information that it generates, it is



necessary to tell the syntactic story. In the case of the subpersonal systems that support and underlie cognition there is always the worry that if the syntactic story is missing one will have attributed to the brain miraculous powers, that is, powers to generate information from information that just couldn't be had by an unintelligent, mechanical system.

It is worthwhile emphasising the beauty of computation from the perspective of the scientific psychologist. In connection with the mind, what is most exciting about computers is not the possibility of building intelligent computers, but rather the actuality that computers are completely stupid. Computers generate meaningful symbols from meaningful symbols and in so doing work things out, solve problems, discover facts, and the like, by entirely mechanical means that involve no exercise of intelligence. This fact about computers raises the possibility that subsystems of the brain engage in computational activity and thus generate information from information by entirely mechanical and unintelligent means. To attribute to such subsystems powers to generate information from information by means of computation in the course of accounting for cognitive capacities is thus not to posit intelligent agents in the brain in any way that threatens circularity or infinite regress.<sup>36</sup> It is this fact about computers and computation that leads scientific psychologists to think of the subpersonal systems that underlie and support cognition as engaging in computational activity. And once they see subpersonal systems as engaging in computational activity, in performing their information-generating tasks, psychologists commit themselves to the task of telling the syntactic story in the course of accounting for cognitive capacities.

In the course of accounting for my capacity to work out my bank balance we saw that it was important to appeal to facts about the world. These facts are such that if they didn't hold certain symbol-manipulating moves would be incapable of generating relevant information. The same point holds of the psychological case. For example, the visual module would not be able to detect such features of surfaces as changes in texture and colour were it not for the fact that there is a systematic relationship between the texture and colour of a surface and the intensity and wavelength of the light that

---

<sup>36</sup> See Dennett (1978).

it reflects. Hence, the psychologist must specify these facts, for a failure to do so would leave it a mystery as to the relevance and success of the computational operations described by her explanation. Indeed, such facts about the world will, in a certain sense, lead the way in the construction of psychological explanations. For facts about the world will provide the psychologist with clues as to how subpersonal systems generate relevant information.

What we have arrived at now is a more or less complete account of the nature of scientific psychological explanation. In summary, here is how that account goes. Scientific psychology aims to account for intentionally-characterised cognitive capacities; in other words to describe how we exercise such capacities. A fundamental assumption is that underlying and supporting such capacities are neurophysiologically-realised subpersonal systems that are capable of generating symbols from symbols by means of computation. These subpersonal systems facilitate cognition because the symbols they manipulate have semantic properties, so that in exercising their symbol-manipulating capacities they generate information from information. In other words, given information as input they generate further information as output. The information generated in this way is precisely the kind that enables the cognitive system as a whole to form the personal-level intentional states that manifest the cognitive capacity in question. Consequently, central to explaining a cognitive capacity is the construction of an account of how the subpersonal system underlying the cognitive capacity in question performs its information-generating task; that is, how it generates the information that it produces as output from the information that it takes as input. As it performs this task by generating information-bearing symbols from information-bearing symbols by means of computation, such an account must include reference to both semantic and syntactic properties. It must describe the formal language employed, the syntactic rules that are applied to the symbols of that language, and the semantic properties of these symbols. A failure to tell the syntactic story will constitute a failure to specify fully how the subpersonal system generates the information that it generates. And a failure to tell the semantic story will leave it a mystery as to the point and relevance of the system's symbol-manipulating activity, for that activity only has a point and relevance

if the symbols manipulated have appropriate semantic properties. Furthermore, those facts about the external world that enable the system to work out the information that it needs from information that it already has by means of the computational operations that it performs must also be specified. Thus the semantic and the syntactic are, as it were, intertwined in scientific psychological explanations. That they are so intertwined is a product of the explanatory ambitions of scientific psychology and the fundamental assumptions of that discipline.

This account of the nature of scientific psychology would appear to imply that scientific psychological explanations are what Cummins calls "property theories" (see fn. 16). In explaining cognitive capacities by appeal to simpler underlying capacities, psychological explanations are not causal explanations of particular events that fit the deductive-nomological model of explanation. However, they can be seen as being causal explanations that effectively appeal to causal laws (or at least to counterfactual-supporting causal generalisations) and for this reason I am hesitant to call them property theories. As we have seen, cognitive capacities (along with the information processing and computational capacities that underlie them) correspond to causal generalisations. Consequently, a system's having such a capacity is just a matter of a certain causal generalisations being true of its input-output behaviour. Moreover, whenever an individual exercises a cognitive capacity, a causal process takes place that is subsumed by the causal generalisation that corresponds to that capacity. And that causal process will be made up of constituent causal processes each of which are exercises of underlying capacities and are thus subsumed by the causal generalisations corresponding to those capacities. (That is why references to underlying information processing and computational capacities can be freely interchanged with references to what goes on causally whenever such capacities are exercised). Effectively then, psychological explanations explain causal generalisations by appealing to the underlying causal generalisations in virtue of which they hold. And they can also be seen as describing what goes on causally whenever the cognitive capacity in question is exercised. For these reasons, scientific psychological explanation is a form of causal explanation.

## 2.7 Fodor's account of psychological explanation

How does Fodor's understanding of the nature of scientific psychology compare with my account? Does my account place any pressure on any of Fodor's views? One difference that we have seen concerns the nature of the RTM that scientific psychology is committed to. According to Fodor, scientific psychology is committed to RTM as a theory of such personal level PAs as beliefs, an assertion that I have rejected. In actual fact, the RTM that underlies scientific psychological theorising is a theory about subpersonal representational states, and not beliefs, desires, and the like.

A consideration of some of Fodor's most explicit pronouncements on the explanatory ambitions of scientific psychology and the nature of the explanations that it endeavours to produce would appear to generate a second point of conflict. Fodor has a tendency to represent scientific psychology as being a sharpened up version of folk psychology which, like the latter, seeks to produce singular causal explanations of PA tokenings and behavioural events that appeal to PAs, and which fit the deductive-nomological model of explanation. The primary difference between the two psychologies is that the scientific version is more thorough, rigorous and research oriented, and so constantly seeks to add to its stock of causal generalisations so as to expand its explanatory powers.<sup>37</sup> Such a view of the nature of scientific psychology and the explanations that it produces would appear to conflict with the account that I have developed in this chapter.

However, the above points notwithstanding, I am hesitant to accuse Fodor of holding a mistaken account of the nature of scientific psychology for there is much that he writes that sits happily with my account. First, he has represented scientific psychology as seeking to explain cognitive capacities by appealing to the simpler underlying capacities and operations of cognitive subsystems.<sup>38</sup>

---

<sup>37</sup> This theme of Fodor's work is most evident in the first chapter of *Psychosemantics*.

<sup>38</sup> In 'The Appeal to Tacit Knowledge in Psychological Explanation' he describes what he labels as "intellectualist" accounts of mental competencies' as answering "how" questions by means of the specification of computer programs. He writes:



Second, Fodor has frequently argued that non-basic laws (and for him all special science laws are non-basic) have implementing mechanisms, and that it is a scientific concern to describe such mechanisms. Here is how he puts the point in *The Elm and the Expert*:

There must be an implementing mechanism for any law of a nonbasic science, and the putative generalizations of psychology are not exceptions. An implementing mechanism is one in virtue of whose operation the satisfaction of a law's antecedent reliably brings about the satisfaction of its consequent. . . . Typically,

---

[A] psychological model in the form of a machine program for simulating the behaviour of an organism ipso facto provides, for each type of behaviour in the repertoire of that organism, a putative answer to the question "how does one produce behaviours of that type?", the form of the answer being a set of specific instructions for producing the behaviour by performing a set of machine operations. Hence to be interested in simulating behaviour is to be interested in a range of "how" questions about behaviour that psychological theories built on the nomological-deductive model are not designed to answer (p. 75).

And as an example of such a psychological explanation he presents the following account of how we tie our shoe laces:

There is a little man who lives in one's head. The little man keeps a library. When one acts upon the intention to tie one's shoes, the little man fetches down a volume entitled *Tying One's Shoes*. The volume says such things as: "Take the left free end of the shoelace in the left hand. Cross the left free end of the shoelace over the right free end of the shoelace . . .," etc.

When the little man reads the instruction 'take the left free end of the shoelace in the left hand', he pushes a button on a control panel. The button is marked 'take the left free end of a shoelace in the left hand'. When depressed, it activates a series of wheels, cogs, levers, and hydraulic mechanisms. As a causal consequence of the functioning of these mechanisms, one's left hand comes to seize the appropriate end of the shoelace. Similarly, *mutatis mutandis*, for the rest of the instructions.

The instructions end with the word 'end'. When the little man reads the word 'end', he returns the book of instructions to his library.

That is the way we tie our shoes. (pp. 63-64).



though not invariably, the mechanisms that implement the laws of a science are specified in the vocabulary of some other, lower level science. Thus, it's a law that water freezes if it is suitably cooled. The mechanism that implements this law involves various changes in the molecular structures of water that suitable cooling reliably induces. (p. 8).

In the case of intentional laws the implementing mechanisms are syntactic and the task of describing these mechanisms falls to scientific psychology. So not all scientific psychological explanation is a matter of explaining PA tokenings or behavioural events by appeal to the PAs that are their causes. My account of scientific psychology can be described as implying that its central task is to describe the mechanisms that implement those intentional generalisations that correspond to our cognitive capacities. When described in these terms a crucial similarity between my account and that of Fodor becomes apparent. However, there is an important point of dissimilarity. Fodor effectively divorces intentional from computational or syntactic psychology; for him, their respective practitioners are involved in distinct explanatory enterprises. I, on the other hand, have emphasised that the intentional and the syntactic/computational are intertwined in explanations of cognitive capacities.

Third, Fodor accepts that scientific psychology recognises the existence of subpersonal representational states and processes. Indeed, appeals to such states and processes figure prominently in his own psychological work. Consider *The Modularity of Mind*. In that book he presents an account of the architecture of the mind that distinguishes between the central cognitive system and various input modules (for example the visual module). Beliefs - and other personal level intentional states - reside within the central cognitive system. The input modules are information-processing systems that present their output to the central cognitive system and thus play a role in belief fixation. However, none of the representational states generated by their activity are beliefs; rather, they are subpersonal states of a subpersonal information-processing system. The processing activity of the input modules is largely unaffected by the subject's beliefs and is therefore cognitively impenetrable. The input

modules are, in Fodor's terminology, informationally encapsulated. Similarly, the information that the input modules have access to in the course of executing their information-processing tasks is limited and domain-specific. For our purposes the important point is that Fodor would accept that scientific psychology is committed to the reality of subpersonal representational states and processes. Moreover, he would see it as a psychological project to determine the workings of the input modules, a project that is part and parcel of that of explaining our cognitive capacities.

Therefore, there are good reasons for not charging Fodor with holding a mistaken or impoverished account of the nature of scientific psychology and psychological explanation despite the fact that he sometimes writes as if that discipline were little more than an extension of folk psychology. However, my account of scientific psychology will play an important role in my evaluation of Fodor's philosophy of mind; many of my criticisms of his most important and provocative claims - along with my contribution to the issues that those claims address - will be motivated by it in one way or another. For, despite all the above, Fodor has a tendency to become fixated on folk psychology and folk psychological explanation when engaged in his philosophical projects, a fact that often has less than happy consequences.

## Chapter 3

# Individualism and the Explanation of Cognitive Capacities

### 3.1 Introduction

How does scientific psychology individuate the representational states that figure in its theories and explanations? Does (or should) it individuate such states individualistically? In this and the following two chapters I will attempt to answer this question.

Arguments within this area (both for and against an individualistic conclusion) tend to fall into one of two categories. On the one hand, there are those of a metaphysical or a priori nature. Such arguments rely heavily on considerations concerning the nature of causation, laws, and scientific explanation in general. On the other hand, there are arguments that rely heavily on quite specific claims about the theoretical commitments, practices, or explanatory ambitions of contemporary scientific psychology.<sup>39</sup> Fodor has produced a series of highly significant arguments for individualism, some of which fall into the first category, and some of which fall into the second. I will discuss these arguments in the course of my reflections. The conclusion that I will attempt to establish in this chapter will be a tentative anti-individualist or externalist one: some of the contents that scientific psychology does, or should, appeal to in the course of explaining our cognitive capacities are not locally supervenient. In arguing for this conclusion I will draw heavily on the account of scientific psychology developed in Chapter 2, and will appeal to a series of pragmatic considerations. However, as I argue in Chapter 4 in connection with David Marr's theory of vision, it is also the case

---

<sup>39</sup> The distinction I have in mind parallels that described by van Gulick (1989) when he writes: 'Some internalist arguments are based on empirical facts about the actual practices and needs of working cognitive psychologists, while others appeal in a more a priori way to very general metaphysical conclusions to support their internalist conclusions' (p. 151).

that some of the contents that figure in the theories and explanations of scientific psychology are locally supervenient. In chapter 5 I will attempt to undermine Fodor's metaphysical arguments for the conclusion that the contents that figure in scientific psychological explanation must be narrow; the nature of causation and scientific explanation are not such as to tell against externalism.

### 3.2 Individualism and folk psychology

Before we can address our target question it is necessary to get clear on just what individualism is. According to Tyler Burge, individualism is the doctrine that:

the mental natures of all a person's or animal's mental states (and events) are such that there is no necessary or deep individuating relation between the individual's being in states of those kinds and the nature of the individual's physical or social environments. (1986, pp. 3-4).

Individualism comes in both a strong and a weak form. According to the strong form an individual's being in a given mental state 'can be *explicated* by reference to states and events of the individual that are specifiable without using intentional vocabulary and without presupposing anything about the individual subject's social or physical environment' (Burge 1986, p. 4). The weaker form is implied by the stronger, being the view that the mental supervenes on the physical so that an individual's mental states 'could not be different from what they are, given the individual's physical, chemical, neural, or functional histories, where these histories are specified non-intentionally and in a way that is independent of physical or social conditions outside of the individual's body' (p. 4).<sup>40</sup>

---

<sup>40</sup> Quite generally A states supervene on B states if an entity's A states could not be other than they in fact are without there being a corresponding difference in its B states. Thus if A states supervene on B states two distinct entities that have just the same B states will have just the same A states. The concept of supervenience is closely associated with the work of Jaegwon Kim. See his 1982, 1984 and 1990.

Most discussions of individualism concentrate on the weaker version of the doctrine. In keeping with this orthodoxy I will take individualism as a thesis about scientific psychology to be the doctrine that that psychology does (or should) individuate psychological states in such a way that an individual's psychological properties supervene on her intrinsic physical properties. Understanding individualism in this way emphasises its connections to two other related claims, these being that psychology is methodologically solipsistic (Putnam, 1975, Fodor, 1980), and that psychology respects the autonomy principle (Stich, 1978). Whatever their differences advocates of these three doctrines all agree that scientific psychology does (or should) individuate psychological states in such a way that the psychological supervenes on the physical.

Before considering scientific psychology it will be useful to review a couple of very famous and much discussed arguments that are widely taken to show that folk psychology does not individuate mental states (or, more specifically, intentional states) individualistically.<sup>41</sup>

The first argument is inspired by Putnam's attempt to establish the conclusion that "meanings ain't in the head".<sup>42</sup> In asserting that "meanings ain't in the head" Putnam is claiming that the meaning of a word on an individual's lips, or his understanding of that word, is not solely determined by his narrow psychological state<sup>43</sup>; rather, it is at least partly determined by the nature of the world external to

---

<sup>41</sup> However, these arguments haven't convinced everyone. A notable exception is Brian Loar (1988). He argues that folk psychology individuates in terms of a notion of content that cannot be identified by what is specified by oblique or de dicto attitude ascription's. One and the same de dicto attitude ascription may be true of two individuals but from this it doesn't follow that the states in question have the same content from the point of view of folk psychology. And two individuals can have states with the same folk psychological content even though the same de dicto attitude ascription cannot be correctly made.

<sup>42</sup> See Putnam (1975).

<sup>43</sup> Putnam defines a narrow psychological state (or a psychological state in the narrow sense) as a psychological state that is permitted by a psychology that makes the assumption of methodological solipsism, that is 'the assumption that no psychological state, properly so called, presupposes the existence of any individual other than the subject to whom that state is ascribed' (p. 220).



the individual. Thus two individuals could be identical in their narrow psychological states yet understand a given word differently due to differences between their home environments. The argument for this conclusion takes the form of a thought experiment which utilises the fact that individuals who are physically identical 'in the sense in which two neckties can be "identical"' (p. 227) will thereby be identical in their narrow psychological states. Here is how the argument goes.

In a distant galaxy there is a planet called Twin Earth that is very much like our own planet. On Twin Earth there is a community of individuals who speak a language very much like English, a community that has a member, call him Oscar<sub>2</sub>, who is a physical duplicate of Oscar, a guy who lives here on Earth. One significant difference between Earth and Twin Earth is that the stuff they call "water" on Twin Earth - the stuff that fills their rivers and lakes, falls as rain, quenches their thirst, and so on - has a physical microstructure that differs from that of the stuff that we call "water"; for it is XYZ rather than H<sub>2</sub>O. In virtue of this difference the English word "water" has a different extension from that of the Twin English word "water"; H<sub>2</sub>O, and only H<sub>2</sub>O, (and therefore not XYZ) falls within the extension of the former whereas XYZ, and only XYZ (and therefore not H<sub>2</sub>O) falls within the extension of the latter. A consequence of this difference in extension is that the English word "water" differs in meaning from its Twin English counterpart. And an upshot of this is that the twins, being fully fledged members of their respective linguistic communities, mean different things by "water" (or understand that word differently) despite their physical (and thus narrow psychological) identity.<sup>44</sup>

Putnam was concerned with linguistic meaning but many have thought that his argument can be extended to generate the conclusion that folk psychology individuates intentional states non-individualistically. Here is one way in which the argument can be extended. Understanding the meaning of a word is an intentional state recognised by folk psychology. The twins, as Putnam has established, understand the word "water" differently. Thus as far as

---

<sup>44</sup> For there to be such a difference in meaning it isn't necessary either that Oscar<sub>1</sub> knows or believes that the stuff he calls "water" is H<sub>2</sub>O or that Oscar<sub>2</sub> knows or believes that the stuff he calls "water" is XYZ.

folk psychology is concerned they are in different intentional states despite their physical identity. Therefore folk psychology individuates non-individualistically. (Pettit and McDowell, 1986. Intro.).

A second way of extending the argument runs thus. We use language to express our thoughts. For example, Oscar uses the sentence "water is wet" to express one of his beliefs, and the sentence "I would like a glass of water" to express one of his desires. Like the sentences that express them, these intentional states have semantic and intentional properties: they are about things in the external world, they have satisfaction conditions, and so on. And just as the semantic properties of the sentences that they express are partly determined by features of the world external to the individual, the semantic properties of these intentional states are partly determined by features of the world external to the individual. For example, because of the nature of his home environment, just as the sentence "water is wet" on Oscar's lips is about water (i.e. H<sub>2</sub>O) and is true if and only if water (i.e. H<sub>2</sub>O) is wet the thought that Oscar expresses with this sentence is about water (i.e. H<sub>2</sub>O) and is true if and only if water (i.e. H<sub>2</sub>O) is wet. Similarly the thought that Oscar<sub>2</sub> expresses with the sentence "water is wet" is about XYZ, and is true if and only if XYZ is wet. In short, the thought that Oscar expresses with the sentence "water is wet" has a different content from that that his twin expresses with the same sentence and thus their respective thoughts would be assigned to different intentional state types by folk psychology. This conclusion generalises to all intentional states that the twins express with sentences containing the word "water"; the thoughts that Oscar so expresses are water thoughts whereas the thoughts that Oscar<sub>2</sub> so expresses are twin water (twater) thoughts. Thus, as far as folk psychology is concerned, the twins diverge in their intentional states.<sup>45</sup>

---

<sup>45</sup> This extension of Putnam's argument does not conflict with what Putnam says despite the fact that a key premise of his argument is the claim that the twins are psychologically identical in virtue of their being physically identical. This is because Putnam means "psychologically identical" to be understood as meaning "psychologically identical in the narrow sense" and in no way commits himself to the idea that folk psychology individuates narrowly or is committed to

The second argument for the claim that folk psychology is non-individualistic was developed by Tyler Burge (1979) and has the advantage of being nowhere near as outlandish as the Putnam-inspired argument. It is also more general in applying not just to thoughts involving natural kind concepts. Burge proceeds by describing an individual who has beliefs involving a concept which he only partially understands. He then describes a counterfactual situation in which the individual is individualistically just as he is in the real world but in which his social environment is significantly different. He then argues that the actual thoughts involving the concept that the individual doesn't fully understand diverge in content from the corresponding counterfactual thoughts and thus that, from the point of view of folk psychology, the actual individual has thoughts that differ in content from those of his counterfactual self.

More specifically, here is one of the cases that Burge presents and takes to be conclusive. An individual has a collection of arthritis beliefs; that is, beliefs 'attributed with content clauses containing "arthritis" in oblique occurrence' (p. 77). Many of these beliefs are true, for example, that he has had arthritis for years, that stiffening joints is a symptom of arthritis, that certain sorts of aches are characteristic of arthritis, and so on. He also has the false belief that he has arthritis in his thigh as he doesn't know that arthritis is, by definition, an inflammation of the joints and that you can't develop it in one's thigh.

Burge next describes a counterfactual situation in which the individual is physically just as he is in the actual situation. In this counterfactual situation linguistic practises are such that the word "arthritis" is typically applied, and defined to apply, to rheumatoid conditions both of the joints and outside the joints. In other words, in the counterfactual situation the word "arthritis" does not mean *arthritis*. Burge argues that due to this fact about linguistic practises in his home environment, the individual in the counterfactual situation 'lacks some - probably all - of the attitudes commonly attributed with content clauses containing "arthritis" in oblique

---

methodological solipsism. Indeed, as Fodor (1980) points out, some of Putnam's comments indicate a hostility to methodological solipsism.

occurrence' (p. 78). 'So the patient's counterfactual contents differ from his actual ones' (p. 79). Burge concludes that:

The upshot of these reflections is that the patient's mental contents differ while his entire physical and non-intentional mental histories, considered in isolation from their social context, remain the same. . . . The differences seem to stem from differences "outside" the patient considered as an isolated physical organism, causal mechanism or seat of consciousness. The difference in his mental contents is attributable to differences in his social environment. . . . such differences are ordinarily taken to spell differences in mental states and events.

In short then, we have a case of physical identity, yet, from the point of view of folk psychology, mental divergence. Hence folk psychology individuates non-individualistically.<sup>46</sup>

However, from the fact that folk psychology is non-individualistic, it doesn't follow that scientific psychology is as well. There might be a mismatch between their respective taxonomies due to a divergence in their theoretical assumptions or their explanatory ambitions. But if scientific psychology individuates in terms of intentional content then for there to be such a mismatch it must attribute a different kind of content to our intentional states. In other words, scientific psychological content must be narrow. Hence the individualist is faced with a dilemma: either she establishes that scientific psychology is not intentional, or she establishes that there is a radical break between folk psychological content on the one hand, and scientific psychological content, on the other. At different points in his career Fodor has seized both horns of this dilemma.

### 3.3 Individualism and computation

In this section I will consider several arguments for individualism that might be labelled "arguments from the Computational Theory of Mind(CTM)". According to such arguments, given the nature of

---

<sup>46</sup> To show that this is not an isolated case Burge presents a whole battery of parallel cases. These cases feature the following terms "sofa", "contract", "brisket", "clavichord" and "red".



computation, scientific psychology's commitment to CTM entails that it is individualistic.

The first of these arguments is due to Fodor and is presented in his classic paper 'Methodological Solipsism Considered as a Research Strategy in Cognitive Psychology'. Methodological solipsism is an approach in psychology that considers the individual in isolation from her environment, attempting to describe her internal mental life in such a way that makes no assumptions about the nature of the external world. In ignoring the individual's environment and making no assumptions about its nature, such a psychology must describe and individuate psychological states in terms of properties that supervene on the individual's internal constitution. Assuming the truth of physicalism, a psychology that endorsed methodological solipsism would therefore individuate psychological states in terms of properties that supervene upon the intrinsic physical properties of the subjects that fell within its domain of inquiry.

Fodor argues that contemporary scientific psychology is methodologically solipsistic and thus, in virtue of its physicalist predilections, individuates in such a way as to respect the local supervenience of the psychological on the physical. His reasoning runs thus. Cognitive psychology is committed to CTM. Now computers only have access to the formal or syntactic properties of the symbols that they manipulate, being blind to their semantic properties (which are determined by features of the world external to the computer). Consequently, cognitive psychology is committed to the idea that mental processes only have access to the formal or syntactic properties of the symbols or representations that they manipulate; in other words it endorses the formality condition. This endorsement of the formality condition entails that the concern of cognitive psychology is to 'study mental processes *qua* formal operations on symbols' (p. 232), and thus that it will ignore the semantic properties of mental representations and states individuating such representations and states in terms of their formal or syntactic properties. In the case of physically realised computers, such properties supervene on their intrinsic physical constitution so that two physically identical computational systems will be formally or syntactically identical. In short, cognitive psychology, in studying and attempting to characterise the



computational processes executed by and in the mind-brain, will consider the individual in isolation from the environment and will individuate mental states and representations formally or syntactically and thus individualistically.

This picture of the nature of cognitive psychology clearly conflicts with that developed in Chapter 2, and so should be rejected. The aim of cognitive psychology is to account for our intentionally characterised cognitive capacities. The subpersonal computational processing that underlies and supports such capacities is able to underlie and support them precisely because of the information processing problems that it solves or the information that it generates. Were it not for the information that the computational processes in my brain generated, I would not have the cognitive capacities that I in fact have. Consequently, a putative explanation of a cognitive capacity that merely characterised the computational processing qua formal symbol manipulation that underlay that capacity would not constitute a full explanation. A complete explanation must characterise the information generated by the computational processing (and thus attribute intentional content to the representations so manipulated) plus those facts about the world that enable the information to be generated in the way that it is. Therefore, cognitive psychology can hardly be methodologically solipsistic in the manner described by Fodor. First, it must attribute intentional content to our mental representations. And second, it must consider and appeal to facts about the environment external to the subject.

A second argument for the conclusion that the commitment of scientific psychology to CTM implies that it individuates individualistically rests on the idea that the intentional properties of mental representations are causally inert or epiphenomenal with respect to the phenomena that scientific psychology seeks to explain. Here is how the argument goes. Suppose an event  $e_1$  causes another event  $e_2$ . Then  $e_1$  will have some property (or properties) in virtue of which it causes  $e_2$ ; such properties are causally responsible for  $e_1$ 's causing  $e_2$ . However, not all of  $e_1$ 's properties will be causally relevant in this way; many will be causally inert or epiphenomenal with respect to  $e_2$ . Consider an example. Edgar's diet causes him to develop a heart condition. His diet has many properties one of which

is the property of consisting of items purchased from Tesco's. It is not in virtue of its having this property that Edgar's diet causes him to develop a heart condition; with respect to that effect this property is causally inert or epiphenomenal. Rather, it is because it has the property of containing high levels of saturated fat that his diet causes him to develop a heart condition. An adequate causal explanation of Edgar's developing a heart condition must specify not just its cause (that is, his diet) but also the causally efficacious property of this cause (that is, its property of containing high levels of saturated fat). This point can be generalised: causal explanations, to be adequate, must specify not just the cause of the effect in question, but also the causally relevant property of that cause.

A consequence of CTM, with its endorsement of the formality condition, so the argument continues, is that the intentional properties of mental states and representations are causally inert with respect to the mental and behavioural effects that such states and representations cause. This is because computers only have access to the syntactic or formal properties of the symbols they manipulate, being blind to whatever intentional properties they have. Therefore, intentional properties should not appear in psychological explanations of behavioural and mental events. Such explanations should specify only the syntactic or formal properties of mental states and representations for it is they that are the causally efficacious properties. This result has bearings on the individuation question. For as intentional properties have no legitimate role in scientific psychological explanations, scientific psychology cannot legitimately individuate mental states in terms of such properties. It has no option other than to individuate mental states in terms of their syntactic properties, properties that are locally supervenient. Therefore scientific psychology, in virtue of its commitment to CTM and subsequent endorsement of the formality condition, is individualistic.

Recently there has been much discussion of the question of the causal efficacy of content.<sup>47</sup> One popular way of responding to the

---

<sup>47</sup> Much of this discussion concerns itself not so much with the consequences of CTM but rather with those of Davidson's anomalous monism. According to Davidson (1970): (i) causation is a relation between particular events; (ii) every token mental event is identical with some token physical event; and (iii) if an event  $e_1$  causes another

threat of epiphenomenalism is to attempt to argue that, contrary to the epiphenomenalist's claim, intentional properties are not causally inert.<sup>48</sup> For the friend of CTM this would involve arguing (as do Block (1990a) and Peacocke (1994)) that computers are sensitive to the semantic properties of the symbols they manipulate. In Chapter 2 I argued that computers are sensitive only to syntactic properties so I do not wish to follow this route. However, I think that all is not lost, for intentional properties are not required to be causally efficacious to have a legitimate role to play in scientific psychological explanation. This is another consequence of the nature of scientific psychology as I have described it. The causal inertness of content might well entail that intentional properties had no legitimate role to play in scientific psychological explanation if the aim of scientific psychology was to produce singular causal explanations of neurophysiologically realised mental events and behaviour that was realised by bodily movements. But the fact that it aims to account for intentionally characterised cognitive capacities makes matters somewhat different. What is crucial about the causal processes that underlie and facilitate perception and cognition is that they generate relevant information. *Ceteris paribus*, if my visual system didn't generate the information

---

event *e*<sub>2</sub>, then *e*<sub>1</sub> must have some property *F*, and *e*<sub>2</sub> some property *G* such that it is a strict law that *F*-events cause *G*-events. As Davidson holds that the only strict laws are physical laws, this doctrine would seem to imply that the intentional properties of mental events are causally inert for whenever a mental event causes an effect it will have some physical property that is sufficient to determine that it produces that effect. For a discussion of this issue see Davidson (1993), Antony (1989) and Heil and Mele (1993).

<sup>48</sup> In 'Making Mind Matter More' Fodor responds to the epiphenomenalist in this way. The version of the epiphenomenalist argument that he considers is the quite general one that all special science properties that are not identical to physical properties, being multiply realisable at the physical level, are epiphenomenal. He argues that a special science property *F* of an event *e*<sub>1</sub> is causally efficacious with respect to an effect event *e*<sub>2</sub> that has special science property *G* if that causal chain is subsumed by the law that *F*'s cause *G*'s (this law need not be strict). He then argues that given the existence of intentional causal laws, intentional properties meet this condition and thus are not epiphenomenal. For a response see Segal and Sober (1990). For some other important contributions to the debate see: Jackson and Pettit (1988), Dretske (1988, 1990), and Yablo (1992).

that it generates from the retinal images that it takes as input, then it would not enable me to perceive the world as I perceive it to be (or to find out about the world by means of vision). Hence an explanation of that capacity cannot ignore the information generated by the visual system's computational activity (and thus the content of the representations it manipulates). A purely syntactic account would leave it a complete mystery as to how the syntactic operations it described were able to support the target capacity and facilitate its exercise; it would leave out what, from the point of view of scientific psychology, was all important about those syntactic operations. All this will be the case even if the visual system is blind to the information carried by the representations that it manipulates; even if content is causally inert. So, in conclusion, the causal inertness of intentional properties does not entail individualism by debarring such properties from having any legitimate role to play in scientific psychological explanation.

A third argument for the conclusion that scientific psychology's commitment to CTM implies that it individuates individualistically has to do with implementation. The argument runs as follows. Higher level laws are implemented by lower level mechanisms. For a higher level law of the form "Fs cause Gs" to be implemented by the lower level mechanism that MFs cause MGs, it must be the case that the instantiation of the property F is sufficient for instantiation of the property MF and that the instantiation of the property MG is sufficient for the instantiation of the property G. Now a scientific psychology committed to CTM holds that intentional laws are implemented by computational mechanisms. Given the general nature of the implementation relation, this implies both that there are computationally sufficient conditions for the instantiation of intentional properties, and that there are intentionally sufficient conditions for the instantiation of computational properties. A consequence of this is that computational mechanisms cannot implement broad intentional laws for it isn't generally the case that there are computationally sufficient conditions for the instantiation of broad intentional properties. (I believe that water is wet, yet being in the computational state that I am in isn't sufficient for having this belief, as my twin is computationally identical to me yet believes that twater is wet). As Fodor puts it, the assumption that intentional



properties are broad 'makes it very hard to see how there *could* be computationally sufficient conditions for their instantiation. How could a process which, like computation, merely *transforms one symbol into another* guarantee the causal relations between symbols and the world upon which . . . the [broad] meanings of symbols depend?' (*The Elm and the Expert*, p.12). In short, computational mechanisms can implement intentional laws only if those laws are narrow so that a scientific psychology that was both intentional and committed to CTM would have to be narrow on pain of inconsistency.<sup>49</sup>

My objection to the above argument is essentially that it misrepresents the nature of the implementation relation; there don't have to be lower level sufficient conditions for the instantiation of higher level properties in order for lower level mechanisms to implement higher level laws. It is certainly true that for computational mechanisms to implement intentional laws there would have to be a close and systematic relationship between computational properties on the one hand, and intentional properties on the other. But that relationship need not be one of coinstantiation (which is effectively how Fodor describes it). In fact, with respect to the case that concerns us, all that the implementation relation requires is the following:

---

<sup>49</sup> In *The Elm and the Expert* Fodor represents this argument as the fundamental motivation for the claim that scientific psychology must be narrow or individualistic. For example, he writes:

The continuing flirtation that a number of philosophers, myself included, have been having with the notion of 'narrow' content over the last decade or so is, perhaps, best understood in this context. It is obscure how externalist intentional laws could be computationally implemented. Very well, then, let there be *another kind* of intentionality - let there be, as one says, 'narrow' content as well as 'broad' content - such that narrow content is ipso facto *not* externalist. And let it be assumed that the content that figures in psychological laws is, in fact, content of this narrow kind. then there could be computationally sufficient conditions for being in the kind of intentional states that psychological laws apply to - viz, for being in narrow intentional states - and everything is fine.



- (i) That intentional state tokens are identical to, or constituted by, computational states.
- (ii) That any subject subsumed by intentional laws is such that intentional differences between its psychological states are reflected in differences at the computational/syntactic level.
- (iii) That any subject subsumed by intentional laws is such that the computational properties that it has are sufficient for the tokening of the intentional properties that it has in the context in which it is embedded. (So that, given its computational states, a subject's intentional states could have been different only if it had inhabited an appropriately different context or environment).

Suppose that an intentional law that Fs cause Gs, and a computational law that MFs cause MGs subsume a subject S. The intentional law will be implemented in S by the computational law if the relationship between the Fs and Gs in S, on the one hand, and the MFs and MGs in S, on the other, satisfies the above three conditions. If this is the case, then whenever S tokens F, that token will be constituted by a token of MF, a token that causes a tokening of MG, something which, in S's circumstances, is sufficient for a tokening of intentional state G.

In effect, what I am arguing is that the implementation relation is such that intentional laws can be implemented by computational mechanisms even if there are no computationally sufficient conditions for the instantiation of the relevant intentional properties; all that is required is that there are context or environment-relative computationally sufficient conditions. This allows the possibility of a computational mechanism implementing an intentional law in me that it doesn't implement in some other subject because that other subject is embedded in an environment suitably different from that in which I am embedded.

Given that a subject typically inhabits the same context or environment for the whole of its life, the intentional laws that subsume it will typically be implemented at the computational level in one and the same way throughout that life. And given that subjects subsumed by one and the same intentional law often share an environment, there is nothing to stop that law from being computationally implemented in them all in just the same way. The

upshot of all this is that computational mechanisms are capable of implementing broad intentional laws but when they do so what enables them to perform this task has a lot to do with the context in which they are embedded.

It might be objected that if matters are as I have described, it is somewhat misleading to talk of computational mechanisms implementing broad intentional laws. For, so the thought goes, it is the computational along with the environmental (rather than the computational alone) that does the implementing. In response, the crucial point is that the broad intentional states are constituted by computational states and not at all by environmental facts. Consequently, the most accurate description of the situation is that intentional laws are implemented by computational mechanisms, but what enables the latter to implement the former are certain facts about the embedding environment. In this connection it is worth pointing out that higher level states, events, and objects can be constituted by lower level states, events, and objects without higher level properties being locally supervenient upon lower level properties. Consider my heart, for example. My heart is firmly located in my chest as it is constituted by a physical object that is located within my chest. Yet my heart is a heart in virtue of facts that lie beyond its outer boundaries, for example, facts about our evolutionary history. Consequently, something could be intrinsically identical to my heart without being a heart and the property of being a heart does not locally supervene upon the intrinsic physical properties of hearts (or of the bodies that contain them, come to that). In other words, my heart is firmly located within my body, despite the fact that it is a heart in virtue of properties that spread out (both in time and space) beyond the outer limits of my body. Similarly, our broad intentional states are quite literally located in our heads in virtue of the spatial location of the lower level states that constitute them, and in spite of the fact that broad intentional properties are not locally supervenient. Hence, the individualist slogan that "psychological states are in the head" is potentially very misleading for in a quite familiar and literal sense most externalists believe that "psychological states are in the head". (Stalnaker, 1989, Davidson, 1987).

In this section I have examined three of the most important and influential arguments for the claim that scientific psychology must be individualistic in virtue of its commitment to CTM. All have been found wanting. I thus conclude that it is consistent with scientific psychology's endorsement of CTM that it is not individualistic.

### **3.4 Practical reasons for avoiding narrow content**

The foregoing considerations emphasise the point that scientific psychology, given its explanatory ambitions and theoretical commitments, must attribute intentional contents to our representational states. But it is consistent with this point that such content is narrow; that scientific psychology attributes locally supervenient contents to our states. In this section I will attempt to outline some serious practical obstacles that face any scientific psychologist who attempts to engage in narrow psychology. The moral I will attempt to draw is that these obstacles are such that the scientific psychologist would be well advised to refrain from going narrow if at all possible. I will also suggest that it is indeed possible to avoid going narrow; that the circumstances in which we find ourselves are such that the advantages of a narrow psychology over a broad one are not so significant as to justify all the additional bother that would be entailed in going narrow.

The first practical problem has to do with the need to construct an adequate notion of narrow content. What exactly is narrow content? What has to be the case for two thought tokens to have the same narrow content? To say that narrow content is locally supervenient and leave it at that is hardly to answer these questions in an illuminating way. A narrow scientific psychology would have to answer such questions and do so by way of the provision of an adequate explication or elucidation of the concept of narrow content that it employed. A failure to do this would have the result that it would be unclear (both to practitioners of scientific psychology and to on-looking outsiders alike) what the aim of scientific psychology was and what claims were being made by a scientific psychologist whenever she presented a narrow explanation or attributed a narrow state to a subject. All this could result in unhealthy confusion within the discipline as psychologists fail to understand the

pronouncements of their fellows and systematically talk at cross purposes. However, developing and explicating a suitable notion of narrow content in such a way as to find favour across the discipline is hardly going to be an easy task. Such a burden would be avoided if scientific psychology were not to deviate from folk psychology in terms of the kinds of intentional properties that it attributed to our psychological states or the concept of content that it employed. Grasping the meaning of the pronouncements of scientific psychologists would be no more problematic a business (both for the practitioners of that discipline and for outside onlookers) than that of understanding everyday folk psychological descriptions and explanations; we could rely quite successfully on our mastery of folk psychological concepts and our facility as folk psychologists. The task of elucidating or explicating the concept of content employed by scientific psychologists would thus become a philosophical one (just as is that of elucidating and explicating the concept of causation). As a general rule, technical concepts that don't have a life and a track record outside of the particular scientific discipline that is their home need to be explicated or elucidated by and for the practitioners of that discipline. The concept of narrow content is such a technical concept, and hence a narrow scientific psychology would be saddled with this burden.

A second practical problem follows on from the first. If scientific psychology is faced with the task of developing and subsequently explicating a concept of narrow content fit to serve its explanatory purposes, then there will always be the possibility of internal conflict within the discipline. Of course conflict and disagreement is part and parcel of any serious scientific activity. But the kind of conflict that I have in mind runs much deeper and is more fundamental than any everyday scientific disagreement. For who is to say that scientific psychologists will agree as to what concept of narrow content is required or as to how the concept of narrow content is to be understood? But if there is no such agreement there will be little possibility of the communication and shared understanding within the discipline that is necessary for its long term health. In short, if scientific psychology attempts to go narrow, then there will be the possibility that its practitioners will come to operate with different concepts of narrow content or divergent understandings of what



narrow content is. And that this would be a likely upshot of scientific psychology's attempting to go narrow (rather than an outside possibility) is suggested by the level of disagreement amongst individualistic philosophers of mind as to how narrow content is to be characterised.

A third problem has to do with the determination and specification of the narrow contents of our thoughts. It is far from easy to determine and specify the narrow content of a subject's thoughts, or establish whether two subjects share a narrow thought. Or at least it is very difficult if we understand narrow content as Fodor does. According to Fodor, narrow content is a function from contexts to broad contents (or truth conditions), so that two thoughts have the same narrow content if and only if they instantiate the same function from contexts to broad contents. Thoughts with the same narrow content that are had by subjects that are embedded in the same context will thereby have the same broad content and hence the same extension. Fodor accepts that, strictly speaking, narrow contents are inexpressible. Any attempt to use a sentence to express the narrow content of a thought will fail, as that sentence will be anchored to a specific context, and thus will have some particular broad content. However, thinks Fodor, this doesn't rule out the possibility of specifying the narrow content of a subject's thought or indicating which narrow thought she tokens. This is because the psychologist can "sneak up" on the narrow content of a thought by mentioning the sentence that has the broad content that the thought in question would have (or its narrow content would determine) if it were embedded in the psychologist's context. For example, I can specify the narrow thought that Oscar2 expresses with the words "water is wet" by mentioning the sentence "water is wet" for, as used by me, that sentence has the same broad content that Oscar2's thought would have if he were embedded in my context. I can do this by saying that Oscar2 has the narrow thought that determines the content *water is wet* in my context. What this implies is that determining the narrow thought of a subject involves determining the broad content that that thought would have were the subject embedded in our context and, consequently, that specifying a subject's narrow thought involves specifying the broad content that it would have were the subject embedded in our context.



This account of narrow content and the means by which the narrow thoughts of subjects are to be determined and specified implies that anyone who attempts to engage in narrow psychology faces some very serious practical obstacles. These are as follows: (i) Fodor seems to overlook the possibility that distinct narrow contents could overlap in their mapping of contexts onto broad contents. Despite being distinct functions the addition function and the multiplication function both map the arguments  $\{2,2\}$  onto the value  $\{4\}$ . Couldn't distinct narrow contents do something similar by mapping one and the same context onto one and the same broad content for one or more (but not all) possible contexts? If this is a possibility then in order to determine the narrow content of a thought, it will be necessary to do more than determine what broad content it would have were it embedded in one's own context. And Fodor's way of sneaking up on the narrow content of a thought isn't going to work, for it will not pick out one particular narrow content. To see this consider the following. Suppose a machine computes a mathematical function. In order to determine which function it computes it wouldn't be enough to determine that for the arguments  $\{2,2\}$  it produces the value  $\{4\}$ , for knowing that wouldn't tell you whether it computed the addition function, the multiplication function or any of the other functions that have the value  $\{4\}$  for the arguments  $\{2,2\}$ . Similarly, to say that the machine computes the value  $\{4\}$  for the arguments  $\{2,2\}$  isn't to specify any one particular function. For just the same reasons, to determine that Oscar2 has a thought that would have the same broad content that the sentence "water is wet" has on my lips were he embedded in my context is not thereby to determine which function that thought instantiates. What I have determined is consistent with his thought's having any of many distinct narrow contents. And the description of Oscar2's thought embedded in the previous sentence does not specify any particular narrow content. Consequently, a narrow psychology would have to do a lot more than Fodor would have us believe in order to determine and specify the narrow thoughts of the subjects under its study. So the question arises: how much more?

(ii) In fact, Fodor's account of narrow content implies that nothing short of determining the various broad contents that a thought would have over a wide range of possible contexts would be enough

to determine the narrow content of that thought. But if that were the case, what hope would a scientific psychology have of ever determining the narrow content of a thought? How, for example, are contexts to be individuated and described? And how, for each such context, is the resultant broad content to be determined and specified? The task looks close to being hopeless.

(iii) I suppose that it is not out of the question that distinct narrow contents just couldn't overlap in the way that mathematical functions do. If it could be established that this were the case, then the arguments presented in (i) and (ii) would be somewhat undermined. However, there would still be major difficulties in determining the narrow contents of the thoughts of individuals who lived in contexts other than our own, difficulties that are obscured by the concentration on twins in the literature. If I know that an individual is my twin, then I can thereby conclude that he has just the same narrow thoughts as I do. And if Fodor's method of determining and specifying the narrow content of a subject's thoughts works, then I can read off the narrow content of my twin's thoughts from the broad contents of my own. But how am I to proceed with respect to an individual who isn't my twin? I can hardly determine the narrow content of its thoughts by determining the broad content of my own. So what am I to do? Determining that individual's broad thoughts wouldn't be enough for that wouldn't tell me whether those thoughts had the same narrow content as mine. And it is far from obvious how I am to work out just what broad content his thoughts would have were he embedded in my context. In short, a narrow psychology is going to have problems in dealing with individuals who inhabit alien contexts and who are not (or who are not known to be) twins of known earthly subjects. And surely the inhabitants of alien contexts who share our narrow psychology are unlikely to be our twins. Perhaps these problems can be overcome, but I reserve the right to be sceptical.

(iv) Of course there is no Twin Earth, and all the real subjects of scientific psychological research inhabit my context or a context closely related to it. This might appear to suggest that in the real world narrow psychology is a practical option; that it is no more difficult to engage in than broad psychology. For all we would need to do to determine the narrow content of a subject's thought would be

to sneak up on it via a determination of its broad content. And we could just as easily translate a broad explanation into a narrow explanation. But if that is all that narrow psychology is, then it is surely something of a scam. Suppose that Edgar, who believes that ferocious dogs savage runners, acquires the new belief that ferocious Fang frequents Brockwell Park. These beliefs causally interact to produce the belief that Brockwell Park is a dangerous place for runners. Fodor's comments would seem to suggest that constructing a narrow description and explanation of Edgar's thought processes would involve determining their broad description and explanation and then effortlessly translating it into something like this: Edgar has the narrow belief that in his/our context has the broad content *ferocious dogs savage runners* and the narrow belief that has the broad content *ferocious Fang frequents Brockwell Park*. These two beliefs interact to cause him to have the narrow belief that has the broad content *Brockwell Park is a dangerous place for runners*. To me that explanation is nothing more than a thinly disguised broad explanation constructed by means of an exercise in broad psychology. A genuine narrow psychology would have to do much more than generate explanations like this; alluding to the narrow contents of broad states is just not enough.

But if a genuine narrow psychology would have to do much more than allude to the narrow contents of broad states, then it isn't going to be so easy to engage in after all, even if, as a matter of empirical fact, all its subjects are locals. Consequently, engaging in genuine narrow psychology, constructing narrow descriptions and explanations of episodes of our mental lives, is going to be a difficult and messy business.

In short, then, narrow psychology faces some significant practical problems. It must develop and explicate an adequate concept of narrow content that finds widespread acceptance within the discipline. Yet in attempting to deal with this burden it runs the very real risk of generating internal dispute and conflict. Even if these problems can be overcome, say by the universal endorsement of Fodor's account of narrow content, it is questionable whether psychologists are ever going to be able to produce any genuine narrow descriptions and explanations. The difficulties presented to narrow psychologists by subjects that share our narrow psychology

but who do not inhabit our local environment and are not our twins will be particularly hard to overcome. And even the narrow psychologist who restricts her attention to her neighbours is hardly going to have an easy time of it.

Of course the existence of such practical obstacles does not in itself entail that scientific psychology does not, nor should not, attempt to go narrow. However, they do suggest that scientific psychology would be well advised to avoid going narrow if at all possible; after all, an impure psychology that we have a chance of making some progress in is to be preferred to a pure one which we have no realistic hope of engaging in successfully. I suppose that the main problem with broad psychology is that it is hopelessly parochial; the broad psychology that we engage in is a psychology of creatures like us living in environments like ours. It is blind to salient psychological similarities between us and our other worldly twins and cousins, and cannot capture the generalisations that subsume us all. But why should we worry about this parochialism when we have yet to discover any other-worldly twins and cousins? If scientific psychologists restrict their attention to us, then broad psychology will work just as well as narrow psychology. After all, if narrow content is related to broad content in the way that Fodor describes, then a broad psychology of subjects who all inhabit the same context will, to all intents and purposes, be nothing other than a locally specific version of narrow psychology. What this suggests is that scientific psychology can, and should, reject the call to go narrow given the problems that it would take on board in going narrow. Given that we have yet to come across any other-worldly twins and cousins, the limitations of a broad psychology would be academic; as a matter of empirical fact, scientific psychology can get away with being broad.

### 3.5 Capturing generalisations

In the previous section I referred to a putative weakness of broad psychology, namely its parochialism. A broad psychology - folk psychology, for instance - is incapable both of recognising and representing some of the significant psychological similarities between us, on the one hand, and our twins and cousins, on the



other, and of capturing some of the intentional generalisations that subsume us all.<sup>50</sup> This would be a major weakness of such a psychology if we had any twins or cousins. But, it might be argued, even if we don't have any twins or cousins, broad psychology is still in trouble. A respectable scientific psychology should be capable of recognising and representing what is psychologically significant about us, and should aim to capture whatever psychological generalisations subsume us. Reflection on our counterfactual fellows indicates that what is psychologically significant about us are our narrow states, and that we are subsumed first and foremost by narrow generalisations. In this section I will attempt to respond to this charge; broad psychology, particularly folk psychology, is not as blind to what binds us together with our twins and cousins as is often thought. Hence we may not, after all, be forced to choose between expediency and generality.

A first point worth noting is that some of the generalisations employed by folk psychology quantify over content. An example is the familiar generalisation that if a person wants it to be the case that *P*, and believes that the best way to bring it about that *P* is by doing *Q*, then, *ceteris paribus*, she will do *Q*. When an individual satisfies the antecedent or the consequent of such a generalisation it will be in virtue of the content of her states. However, individuals whose states have quite different contents can be subsumed by such a generalisation. Hence folk psychology can capture some of the generalisations that subsume both us and our twins and cousins, and can recognise some psychologically significant similarities between us that are not a matter of sharing broad contents.

Yet not all generalisations and similarities can be captured in this way. Any serious psychology needs to recognise and employ generalisations that are less abstract, that appeal to specific psychological states. This is because some states will have quite distinctive causal powers. For example, we would expect Edgar's belief that he was abused as a child, or that he is systematically hated by his fellows, to have causal ramifications that find no parallel in those of his beliefs that giraffes have long necks or that his mother was born in Nova Scotia. Hence we will need to appeal to content-

---

<sup>50</sup> An expression of the complaint that broad psychology is unable to capture important generalisations can be found in Block (1991).



specific generalisations to describe and account for some elements of our psychology. This gives rise to the worry that a broad psychology, in attempting to capture such content-specific generalisations, will inevitably go too far and miss some of the generalisations that subsume both us and our twins and cousins.

But perhaps all is not lost for folk psychology, for there would appear to be a way of capturing similarities between intentional states that diverge in their broad content that folk psychologists frequently and routinely employ. Edgar has an indexical belief that he might express with the words "once when I was out running I was savaged by a ferocious dog". Waldo suffered a similar experience and holds a corresponding belief. These two beliefs differ in broad content, as Edgar's is about Edgar and a past episode in his own life, whereas Waldo's is about Waldo and a past episode in his life. Yet any folk psychologist will tell you that there is a significant similarity between the respective beliefs of these two individuals: they both believe that they have been savaged by a ferocious dog. So here we have a similarity between broad-content-divergent beliefs that is not a matter of their having the same narrow content. Moreover, this similarity may well show up in the subsequent mental and behavioural lives of Edgar and Waldo. Both might become nervous when they see a dog whilst out running, or harbour a long-standing dislike of dogs, or something else along those lines. Hence there may well be a generalisation such as this: anyone who believes that they have been savaged by a dog in the past becomes nervous whenever they are confronted with an unfamiliar dog, *ceteris paribus*. As Edgar and Waldo run through the park together they see Fang in the distance. Thus they both satisfy the antecedent of this generalisation in virtue of their respective broad states, and they subsequently become nervous. Here one and the same highly specific intentional generalisation governs the mental life of two individuals with distinct broad states, and explains why they both become nervous. But the generalisation in question is not narrow.

However, it might be objected, an ability to capture generalisations where thoughts with an indexical component are involved gives us no indication as to how folk psychology could deal with those problem cases where the thoughts involved are not indexical. But another line of defence is suggested by the fact that the embedded *that*

clauses of belief - and other attitude - ascribing sentences often do not serve to specify the content of a belief of the subject in question but, rather, some fact about her that holds in virtue of the content of the beliefs that she has. To see this, first of all consider the practice of using sentences with embedded *that* clauses to describe the claims or assertions made by an individual.

In the course of our everyday life we frequently use language to make claims or assertions about the world. We also use language to describe and report the claims and assertions made by ourselves and our fellows, as when I say "Edgar claimed (/asserted/stated/said) that Fang is dead". What is the function of the embedded *that* clause in such sentences? Typically such clauses do not serve to indicate the specific sentence used to make the assertion in question or the content of that sentence. Rather, they serve to indicate some aspect of the commitment as to how the world is that the subject makes in using the sentence that she utters to make an assertion. When an individual uses a sentence to make a claim or assertion she thereby makes a whole body of commitments as to how the world is; in other words, she commits herself to the truth of a whole collection of distinct propositions. Just what she commits herself to will depend an awful lot on the content of the sentence that she utters, but other factors sometimes play a role. I have in mind three distinct kinds of case.

(i) Waldo sincerely says "Fang has been assassinated". He thereby commits himself to Fang's having been assassinated or, if you prefer, to the truth of the proposition that *Fang has been assassinated*. Thus he can correctly and legitimately be described as having claimed (/asserted/said/stated) that Fang has been assassinated. But he also commits himself to the truth of a whole load of other propositions, the proposition that *Fang is dead*, for example. Thus Waldo can just as legitimately be described as having claimed that Fang is dead and that is how we would describe him in certain contexts when certain interests were operative. Notice that when I say "Waldo claimed that Fang is dead" I do not seek to specify the sentence by means of which he made his assertion or the content of that sentence; the sentence embedded in the *that* clause of my description is not the sentence that Waldo uttered, nor does it have the same content as that sentence. Nor do I seek to specify a sentence that Waldo is disposed to

utter, or the content of such a sentence. When I utter the sentence "Waldo claimed that Fang is dead", I seek to highlight some aspect of the commitment that he makes (or one of the propositions the truth of which he commits himself to) in using the sentence "Fang has been assassinated" to make an assertion; I do not seek to announce my beliefs as to what he is disposed to say. That Waldo commits himself to the truth of the proposition that *Fang is dead* clearly has to do with the content of the sentence that he utters; the content of that sentence logically implies that Fang is dead and that is why he can be described as having claimed "that Fang is dead".

Which aspect of the commitment made by an individual in making an assertion that we seek to highlight will be a context-dependent and interest-relative matter. If someone comes up to me and says "I've heard that Fang is dead. Have you heard any such news?" I may well answer by saying "Yes, Waldo told me that Fang is dead". But if I am asked "Do you know anything about the cause of Fang's death?" I may well reply "Yes, Waldo told me that Fang has been assassinated" (in this context replying "Waldo told me that Fang is dead" is clearly a useless and inappropriate answer). Similarly, whether I describe Waldo and Edgar as having made the same claim in uttering their respective sentences will be a context-dependent and interest-relative matter. In certain circumstances I will want to highlight their shared commitments, and will describe them both as having asserted that Fang is dead. In other circumstances I will want to highlight the fact that Waldo goes beyond Edgar in committing himself to the truth of the proposition that *Fang has been assassinated*, a proposition the truth of which Waldo does not commit himself to.

(ii) Sometimes the nature of the commitment made by an individual in using a sentence to make an assertion will be affected by the doxastic surroundings of that speech act. Suppose that Edgar believes that Fang is a ferocious dog. Whilst running through the park he sees Fang and utters the sentence "Fang is in the park". In virtue of the content of that sentence and what he believes about Fang, Edgar commits himself to the truth of the proposition that *there is a ferocious dog in the park* in making his assertion. Consequently, he can legitimately be described as having made the claim that there is a ferocious dog in the park. Had Edgar not had any beliefs concerning the ferocity of Fang, then his assertion couldn't be so described. In

general, the nature of the commitment that an individual makes in using a sentence to make an assertion can be affected by what she believes and this fact is often reflected in how we describe the assertions of our fellows, for we often take into account what we know about what they believe.

(iii) In another kind of case the embedded sentence in the *that clause* of a sentence of the form "she claimed (asserted/stated/said) that . . ." serves to summarise or capture the essence of a whole collection of distinct assertion-making utterances. An example of such a case is that where we describe a weather forecaster's detailed comments about tomorrow's weather in various parts of the country by means of the sentence "she said it will be lousy right throughout the country all day tomorrow". Here, neither is the content of a particular sentence used by the weather forecaster to make an assertion about the weather conditions somewhere, nor the commitment made in so using such a sentence, being specified. Rather, a global commitment made in making a whole series of distinct and less general commitments is specified; a global commitment that is logically implied by the body of less general commitments.

In short, then, what these three distinct kinds of cases indicate is that when we describe the nature of an assertion made by an individual in uttering a sentence (or a collection of sentences), what we seek to specify is typically not the sentence uttered or its content but, rather, some aspect of the commitment about the world that the individual makes in uttering the sentence(s) in question. The content of the sentence(s) will play an important role in determining the precise nature of that commitment, but other factors, such as the individual's related beliefs, will also play an important role. And which aspect of the commitment we choose to focus on will depend upon the context and our operative interests.

My contention is that parallel points hold of belief-attributing sentences. Our aim in using sentences of the form "she believes that . . ." is typically not to express or indicate the content of a particular belief. Rather, it is to indicate some aspect of the commitment that an individual makes in holding a particular belief or collection of beliefs. The three distinct cases described above all have analogues. Suppose that Waldo has a belief the content of which is *Fang has been assassinated*. In having that belief he commits himself to the



truth of the proposition that *Fang has been assassinated*, and, in virtue of the logical implications of that proposition, to the truth of the proposition that *Fang is dead*. Consequently, he can legitimately be described as believing that Fang has been assassinated and as believing that Fang is dead. Which way we will describe him in practice will be a context-dependent and interest-relative matter. In situations where we want to stress the similarity between Waldo and Edgar (who has a belief the content of which is *Fang is dead*), the latter description will be employed.

The analogue of (ii) is the case where Edgar, who has a long-standing belief with the content *Fang is a ferocious dog*, forms a belief with the content *Fang is in the park*. In forming this latter belief, in virtue of its doxastic surroundings, he commits himself to the truth of the proposition that there is a ferocious dog in the park. He so commits himself regardless of whether he forms a belief with the content *there is a ferocious dog in the park*. Consequently, the very belief that Edgar forms when he sees Fang can be described as a belief that Fang is in the park or, alternatively, as a belief that there is a ferocious dog in the park. That his belief can be legitimately described in the latter way is a product of the fact that the commitments an individual makes in coming to hold a particular belief can be affected by that belief's doxastic surroundings.

The analogue of (iii) is the kind of case where in saying "she believes that . . ." we are concerned not with a particular or specific belief state of the subject in question but rather with a more global state. In this kind of case we attempt to summarise or capture the essence of a whole series of distinct beliefs by specifying a general commitment that the subject makes in virtue of holding that collection of beliefs. An example is the case where I describe the weather forecaster as believing that the weather will be lousy all over the country all day tomorrow on the basis of learning that she believes that it will rain in the North, believes that it will hail in the Midlands, believes that it will snow in the South, and so on.

It might be objected that I have misrepresented what is going on in these kinds of cases, that in actual fact what we are doing is ascribing to the subject a dispositional belief with a specific content. My reply is that there are no more grounds for the claim that these belief-ascribing sentences ascribe dispositional beliefs than there is for the



claim that the sentences that we use to describe what our fellows have asserted primarily serve to attribute to them a disposition to utter the sentence embedded in the *that clause*. When, in response to hearing Waldo say "Fang has been assassinated" I say "Waldo claimed that Fang is dead", I am talking about what he claimed in uttering the sentence that he uttered; I am specifying an important aspect of what he committed himself to in executing that speech act. I see no reason to believe that matters are appreciably different in the case where I say "Waldo believes that Fang is dead" in response to hearing his utterance; I am describing some aspect of the very belief that underlies his speech act (a belief that presumably has the content *Fang has been assassinated*).

None of this is to deny that beliefs have specific and determinate contents, that they are individuated in terms of their content, or that we are sometimes concerned with determining and specifying that content. It's just that a lot of the time our primary concern is not so much with the content of the beliefs of our fellows but with the commitments as to how the world is that they make in having the beliefs that they have. There are very good reasons for this. First, what an individual is committed to can be relied on to make some difference to her subsequent mental and behavioural life, and so a knowledge of such commitments can be of significant explanatory and predictive use. Second, it is much more common for people to share commitments about how the world is than it is for them to share beliefs with just the same content. Hence if we want to capture psychological similarities between people it is useful to focus our attention on their commitments as to how the world is, rather than on the contents of their beliefs. And thirdly, it is often very difficult to determine the precise content of the beliefs of our fellows, not least because we can be so inarticulate when it comes to describing our mental life. The salient aspects of our commitments as to how the world is are typically much more easily discernible.

All of this suggests a defence of folk psychology against the charge that it is hopelessly parochial, that it is blind to significant similarities between ourselves on the one hand, and our twins and cousins on the other and that it is incapable of capturing generalisations that subsume us all. Perhaps my beliefs diverge in content from those of my twins and my cousins, but that doesn't

debar a coincidence in some of the commitments that we make in believing what we believe. And although we might not be subsumed by the same broad content generalisations, we might be subsumed by generalisations that (primarily) invoke commitments as to how the world is. For example, Oscar has a belief the content of which is that *water quenches thirst*. Oscar2 has no such belief, but he does have a belief the content of which is that *twater quenches thirst*. In virtue of the doxastic surroundings of their respective beliefs, they are both committed to the truth of the proposition that the colourless liquid that comes out of taps is thirst-quenching.<sup>51</sup> Consequently, they might both be subsumed by the generalisation that anyone who is thirsty and believes that the stuff that comes out of taps is thirst-quenching will take a drink from the nearest tap, *ceteris paribus*.

For reasons such as the above I am far from convinced that a broad psychology (such as folk psychology, for example) is hopelessly parochial. Broad psychology may well have devices at its disposal to enable it to recognise and represent similarities between individuals whose states are divergent in their broad content. So perhaps a scientific psychology need not describe and individuate psychological states in terms of their narrow content, or appeal to narrow content generalisations in order to be a psychology sufficiently general to cover both us and our other-worldly twins and cousins.

### 3.6 Explaining cognitive capacities

In earlier sections I appealed to the account of scientific psychology developed in Chapter 2 to undermine several arguments for

---

<sup>51</sup> It might be objected that the twins are not committed to the truth of one and the same proposition due to the fact that they are related to different taps. (The idea is that Oscar is committed to the truth of the proposition that the colourless liquid that comes out of Earthly taps is thirst-quenching whilst Oscar2 is committed to the truth of the proposition that the colourless liquid that comes out of Twearthly taps is thirst-quenching). However, this point can be conceded without any damage being done. For the twins can be bound together by a combination of an appeal to their commitments and an application of the technique employed to lump together broad-content-distinct indexical beliefs: they are both committed to its being the case that the colourless liquid that comes out of their local taps is thirst-quenching.

individualism. In this section I will attempt to bring that account to bear to establish a more positive conclusion, namely that some of the intentional properties that scientific psychology appeals to in explaining our cognitive capacities are not locally supervenient.

Scientific psychology is not sharply cut off from folk psychology, for it attempts to account for facts about us that are very familiar from the folk psychological perspective; facts that we frequently rely upon in going about our folk psychological business. Such facts are that we are able to recognise our friends, classify objects, understand sentences of natural language, recall past events, and so on. These cognitive capacities are broad in the respect that their exercise manifests itself in the tokening of a broad intentional state. For example, it is central to the identity of my capacity to recognise my friends that its exercise manifests itself in the tokening of a belief that represents the identity of the friend before me. Such properties as that of representing the individual before one as being Edgar are clearly not locally supervenient; I could believe that the individual before me was Edgar without my twin (who has had no contact with Edgar and has never even heard of him) believing that the individual before him was Edgar.

In order to explain such broad capacities, or describe how we exercise them, the scientific psychologist must attribute broad intentional properties to some of our subpersonal representational states, or make it explicit that those states have such properties. Suppose that I come across Fang when out running and, exercising my capacity to recognise familiar individuals when confronted by them, I form the belief that Fang is before me. How did I do this? How did I exercise that capacity? The cardboard cut-out version of the answer runs thus. My visual system generates a representation of Fang's observable properties (his shape, size, colour, etc.) which is fed to the recognition module. The recognition module has access to a whole battery of representations stored in long-term memory that represent various properties (including the visual appearance) of the many individuals of my acquaintance. Amongst these representations is one that represents Fang's appearance, that carries information about his shape, size, colour, and so on. The recognition module compares its input representation with those stored in long-term memory until it finds a match; that is, until it comes across a

representation in long-term memory that represents a certain individual as having just those observable properties that the visual system represents the distal cause as having. In this case the match is made with the representation that carries information about Fang. This results in the generation of a representation reporting this match which is fed to the central cognitive system eventuating in the formation of a belief that Fang is present.

Let us focus our attention on the representation stored in long-term memory, a representation that plays the fundamental role in the recognition process. It carries information about Fang, in particular about his visual appearance or observable properties. Its being about Fang, its having the content that Fang has such and such an appearance, is fundamental in enabling it to play any role in the process of recognising Fang. If it carried information about the appearance of some other dog, then I would end up believing that that dog was before me; it would be of no use in enabling me to recognise Fang. Consequently, an explanation of how I recognise Fang must emphasise the point that this representation represents Fang's appearance or carries information about Fang. A putative explanation that ignored the reference of this representation would leave it a complete mystery as to how it facilitated my recognition of Fang.<sup>52</sup> But the property of carrying information about Fang's observable properties (or that of representing his visual appearance as being a certain way) is not locally supervenient. The corresponding representation in my twin carries information not about Fang but about some other dog (how could it carry information about Fang given that he has never come across that dog or even heard about him). If my twin came across Fang he would end up forming a false

---

<sup>52</sup> An analogous point can be made about my capacity to work out my current bank balance. I do this by manipulating symbols that carry information about previous states of my bank account and episodes in my banking history. It is of fundamental importance that these symbols are about my banking history for it is because of their reference that I am able reliably to generate from them information about my current balance. Were they about previous states of Edgar's bank account and episodes in his banking history then I would not be able to work out my current bank balance from them. Consequently, an explanation of my capacity must specify such facts about the representations that I manipulate, otherwise it will leave it a complete mystery as to why the processes that I execute work.



belief as to the identity of the beast before him. What would explain his mistake would be the fact that Fang has just the same appearance as his representation of this other dog represents that dog as having. Therefore, a scientific psychology that attempts to explain my capacity to recognise Fang must attribute to the subpersonal representational states that it appeals to intentional properties that are not locally supervenient.

My capacity to recognise Fang is not peculiar in this respect. Underlying and explaining many of our cognitive capacities are subpersonal representational states that have significant intentional properties that are not locally supervenient. What enables me to recognise and classify objects (as well as individuals) are representations that carry information about the specific kinds of objects that populate my environment. What enables me to recall past events in my life are representations that carry information about those events. And so on.

It might be objected that my argument doesn't tell against individualism, as I have not established that the reference of the states that I have described plays any role in determining their psychological nature. In greater detail, here is the argument that I have in mind. Consider the capacity to perceive the world by means of vision. Whenever I exercise this capacity I token a visual state that represents such properties as the shape, size, colour, motion, etc. of whatever object is currently impinging on my visual apparatus. Such visual states are about the objects impinging upon me; they represent those objects as being a certain way, something that they would have to do to count as manifestations of the capacity to perceive the world by means of vision. Yet the fact that such states need to have a quite specific reference to count as a manifestation of the capacity in question doesn't entail that reference plays any role in the individuation of such states. Suppose that I am confronted by a square-shaped object in response to which I token a visual state that represents that object as being square-shaped. The identity of the object in question doesn't enter into the content of my visual state; my state would have just the same content (and would thus belong to the same psychological type) no matter which particular square-shaped object it represented. And the corresponding state of my twin would have just the same content as my state even though he was



being stimulated by a different square-shaped object (or if his state did have a different content than mine this would not be due to our respective states diverging in their reference). What this raises is the possibility that though the representations implicated in object-recognition have quite specific references (as they would have to have in order to support the capacity of object recognition) those references do not enter into their content. It is one thing for a representation in me to represent Fang's characteristics (or carry information about Fang) whilst the corresponding representation in my twin represents the characteristics of some other dog (or carries information about some other dog), but it is quite another for our respective representations to diverge in their content in virtue of this difference.

My response to this argument runs thus. In order for me to have the capacity to recognise Fang it has to be the case that (typically) whenever I am confronted by Fang a causal process is set off that eventuates in my forming a belief to the effect that Fang is before me. This belief isn't just about Fang in the way in which a Fang-caused visual state is about Fang. Rather it represents my distal stimulus as being Fang; it has a content something along the lines of *Fang is before me*. Now the corresponding belief of my twin clearly does not have this content. It is not true of my twin that if he were confronted by Fang he would come to believe that Fang was before him. In other words, he does not have the capacity to recognise Fang. Similarly, the content of the representation underlying my capacity to recognise Fang doesn't just carry information about Fang or have Fang as its reference. Rather it represents Fang as having certain characteristics; it has the content *Fang is such and such*. It is this fact about that representation that enables it to support my capacity to recognise Fang as Fang. If my recognitional module didn't "know" what Fang looked like (if it only believed that there existed a dog of unknown identity that had certain visible characteristics) then it would not be able to help me recognise Fang when confronted by that dog. My twin's recognitional module doesn't "know" what Fang looks like. If it did, he too would be able to recognise Fang. But it does know what some other dog (a dog with whom I am unacquainted) looks like. The point generalises: reference enters into the content of all of the representations stored in long term memory that carry information

about particular objects, individuals, events, types of substance, and the like. If this were not the case we would not have the recognitional and classificatory capacities that we in fact have. And at least some of these contents are not locally supervenient; for example the content of those representations that represent the appearance of individuals of our acquaintance, and that that (in inhabitants of a watery world) represents the appearance of water.<sup>53</sup> Consequently, in order to account for our familiar, everyday, recognitional and classificatory capacities, scientific psychology must attribute to some of our representational states contents that are not locally supervenient.

Given that some of the cognitive capacities that scientific psychology attempts to explain are broad, going narrow is not an option for the scientific psychologist. A narrow psychology could, at best, only explain narrow capacities. But, it might be objected, such narrow accounts of narrow capacities could be supplemented with the details of the subject's environmental embedding so as to generate an explanation of the corresponding broad capacity. The objections to this tactic should by now be familiar. First, won't such an explanation in effect be a thinly-disguised broad explanation, and not a genuine exercise in narrow psychology? If (as Fodor thinks) narrow content determines broad content relative to context, then it is difficult to avoid the conclusion that to attribute a narrow state to an individual and then specify the details of her environmental embedding is in effect to attribute to her a particular broad state (and to do so in a particularly messy and disingenuous way). Second, it will be far from easy to determine and specify the details of our environmental embedding. A far more realistic and practical option is to, so to speak, build the details of the environmental embedding into the intentional-property attributions. After all, in doing this we will not be making any claims that are false (broad properties are just as real

---

<sup>53</sup> One of my capacities is that of being able to distinguish between water and other liquids, and being able to recognise water as such whenever I come across a sample of that substance. When I exercise this capacity I form a belief that the stuff before me is water. And what underlies and explains this capacity is a subpersonal representation that represents the observable properties of water, a representation that has the content *water looks thus and so*.

as narrow ones), nor will we be any less able to satisfy our explanatory ambitions.

But why, it might be asked, must scientific psychology attempt to explain the broad capacities that I have described? Why can it not pursue a different explanatory agenda, namely that of explaining the corresponding narrow capacities, capacities whose exercise manifests itself in the tokening of a narrow state? My reply is to say that even if scientific psychology could (in principle) set itself the task of explaining such narrow capacities, what motivation is there for adopting this goal? We really have the capacities that scientific psychology seeks to explain, and those capacities are inevitably going to be of particular interest to us given our immersion in folk psychology. In general, science has to speak to our desires and interests (that is, answer the questions that strike us as important and interesting), otherwise there is little point in our engaging in it (not least because of the cost to the tax-payer of scientific research).

Our twins and cousins will have many of the broad capacities that we do.<sup>54</sup> For example, they will be able to recognise their friends, categorise objects, recall past events in their lives, and so on. And in an important respect they will exercise these capacities in just the way that we exercise ours. Consequently, the explanation of their capacities will be just the same as that of ours. This doesn't contradict anything that I have said for the following reason. Suppose that you work out your bank balance in just the same way that I work out mine. For this to be the case, it is crucial that the symbols that you manipulate represent facts about episodes in your banking history and not facts about episodes in my banking history. In other words, for one and the same explanation to account for both your and my capacity to work out our respective bank balances, it is necessary that the symbols that we manipulate differ in their intentional properties. The same point applies to the explanation of cognitive capacities. Such explanations are pitched at a level of generality that enables them to account for the capacities of individuals whose states diverge

---

<sup>54</sup> Of course my twin will not have the capacity to recognise Fang or the capacity to recognise water, as he doesn't have the concepts of *Fang* or *water*. But those capacities are instances of more general capacities that my twin certainly will have, namely those of being able to recognise familiar individuals, and classify familiar types of stuff.

in their intentional properties. Yet these explanations will make it clear that the representations manipulated by the processes that they describe have (and need to have) intentional properties that represent the characteristics of quite specific individuals, types of objects, historical events, and the like. And these are just the sort of properties that are not locally supervenient. Therefore, scientific psychology doesn't produce parochial explanations; its explanations account for the cognitive capacities of our twins just as much as they account for our cognitive capacities. This is so notwithstanding the fact that scientific psychology attributes to the representational states underlying our capacities intentional properties that are not locally supervenient. Thus it would be misguided to go narrow in response to a desire for generality.

### **3.7 Fodor's account of narrow content**

In section 3.4 I outlined some problems with Fodor's account of narrow content. I argued that his manner of determining and describing the narrow content of our psychological states won't work, and that narrow psychology as he characterises it is a thinly-disguised broad psychology. In this penultimate section I will add to my objections.

The first objection runs thus. Fodor's account is very vague and unclear. He says that narrow content is a function from contexts to broad contents, and that narrow content determines broad content relative to context, so that, for example, our water thoughts must be about water, given their narrow content and the fact that they are anchored to water. Yet he doesn't tell us anything about how contexts are to be individuated and described. Neither does he tell us what is involved in a thought or a subject's being embedded in a particular context, or how we are to understand the anchoring relationship. For example, if I were transported to Twin Earth my water thoughts would not become twater thoughts, so I cannot be embedded in my Twin's context; in this case physical presence is not enough. So what would be the difference between me and my twin in virtue of which he was embedded in his context but I was not? In the absence of answers to such questions Fodor's account is hopelessly vague and incomplete.



This criticism is not as petty and as pedantic as first might appear. Fodor presents his account of narrow content in the context of defending intentional psychology against the charge that it is unscientific. He accepts that folk psychology won't do as legitimate science (in virtue of its not employing an individualistic taxonomy), but argues that this doesn't rule out the prospects of a scientifically-respectable intentional psychology. For such a psychology can employ a notion of narrow content. But given the widespread suspicion that the whole notion of narrow content is incoherent he will not have succeeded in vindicating intentional psychology's claim to be a legitimate science until he has developed a clear and complete account of narrow content; in the circumstances we have a right to remain sceptical if all he does is make some gestures towards such an account.

A second objection is that a psychology that employed Fodor's notion of narrow content would be, in certain respects, too course-grained. It would imply that certain intuitively quite distinct thoughts belong to the same intentional type. I believe that Fang is ferocious. I also believe that Gnasher is ferocious. These beliefs have different broad contents. But it would appear that they have just the same narrow content as they instantiate the same function from contexts to broad content. Fang might have been called "Gnasher"; there is a possible world or context where Fang has that name. Similarly, there is a possible world or context where I am individualistically just as I am but where my "Gnasher" thoughts are about Fang, where they are anchored to that dog. In other words, in the context where they are anchored to Gnasher, my "Gnasher" thoughts are about Gnasher, and in the context where they are anchored to Fang they are about Fang. But exactly the same is true of my "Fang" thoughts; were they anchored to Gnasher (as they could have been with me being individualistically just as I am now) they would be about that dog. In other words, these thoughts instantiate just the same function from contexts to broad context and thus have identical narrow contents. The fact that they differ in broad content, is a product of their being embedded in different contexts rather than their having divergent narrow contents. But is it acceptable for a scientific psychology to assign my belief that Fang is ferocious to the same intentional type as my belief that Gnasher is ferocious? Surely



scientific psychology should distinguish between these beliefs; after all, they do have quite different causal powers.<sup>55</sup>

My final objection is related to the second, as it again accuses Fodor of lumping together states that a scientific psychology should distinguish. The belief that I express with the words "water is wet" has the same broad content as that that I express with the words "H<sub>2</sub>O is wet". Intuitively, though, there is an important semantic difference between these beliefs; they differ in sense, or something like that. However, they do not differ in narrow content. Were my "water" thoughts anchored to XYZ (as my twin's "water" thoughts are) they would be about XYZ. In other words, their narrow content is such as to entail that they are about water in a watery context, twater in a twatery context, and so on. But exactly the same is true of my "H<sub>2</sub>O" thoughts for were I embedded in my twins context they would be about XYZ. (If that sounds implausible ask yourself what my twin's "H<sub>2</sub>O" thoughts are about. It's difficult to see how they could be about anything but XYZ). Consequently, any thought that I express with the words "water is F" will have just the same narrow content as the thought that I express with the words "H<sub>2</sub>O is F", and hence, according to Fodor, belong to the same scientific psychological type. But surely a scientific psychology should distinguish between these thoughts.<sup>56</sup> Once again, these thoughts have quite distinctive causal powers.

I therefore conclude that there are serious problems with Fodor's account of narrow content.

---

<sup>55</sup> For example, my belief that Fang is ferocious causes me to hide when I see Fang whilst out running, but not when I see Gnasher. And my belief that Gnasher is ferocious causes me to hide when I see Gnasher whilst out running, but not when I see Fang.

<sup>56</sup> In his recent writings Fodor has argued that scientific psychology individuates partly in terms of syntax. This implies that it would assign my "water" thoughts to different types than their corresponding "H<sub>2</sub>O" thoughts because of differences between their respective syntactic properties. In Chapter 7 I attempt to undermine this claim.

### 3.8 Conclusion

In this chapter my reflections have taken a decidedly anti-individualist line. By way of summary, my key points can be described in the following manner. The endorsement by scientific psychology of the computational theory of mind does not commit it to individuating psychological states individualistically, several key arguments to the contrary notwithstanding. There are practical problems associated with any attempt to engage in a narrow psychology, problems that may well outweigh any benefits that such a psychology could bring us; whether the possible benefits are outweighed by the drawbacks may well be contingent upon such factors as whether we have any other-worldly twins or cousins. Moreover, the advantages of narrow over broad psychology have often been exaggerated; for example, it is far from clear that scientific psychology has to go narrow in order to recognise the similarities between us and our twins and cousins and capture the generalisations that subsume us all. When addressing the question of how scientific psychology individuates psychological states it is of paramount importance to bear in mind the actual explanatory ambitions of its practitioners. As a matter of fact, or so I contend, in keeping with the continuity of scientific psychology with folk psychology, they are often concerned with accounting for broad cognitive capacities and, as a result, sometimes appeal to intentional properties that are not locally supervenient. That scientific psychologist's should seek to account for capacities that are salient from the folk psychological perspective, and particularly interesting to those who adopt that perspective, should come as no surprise; at the end of the day what we all want is for science to answer those questions and account for those facts about the world that seem particularly interesting and important to us.

I recognise that none of this conclusively vindicates the thesis that scientific psychology is externalist. This is because individualism is a thesis about how psychological states are (or should be) individuated. The fact that scientific psychology is (or could get away with being) broad implies only that some of the intentional properties that it recognises and appeals to are not locally supervenient; it doesn't

imply that those broad intentional properties are the properties that are individuating in scientific psychology. I have little by way of argument for the claim that broad intentional properties are individuating. However, it is far from obvious to me that there is any determinate fact of the matter as to which of the many properties that scientific psychologists appeal to in their theories and explanations are individuating and which are not. It is often taken for granted that scientific psychology operates with a clear-cut, hard and fast taxonomy, so that there is always a determinate once and for all answer to the question of whether two psychological state tokens belong to the same psychological type. I want to resist this assumption. Suppose two psychological state tokens share certain properties recognised by scientific psychology. It may well be the case that there is a range of legitimate scientific psychological interests, concerns and purposes that are such that when some of them are operative the similarities between the state tokens will be particularly salient and relevant, whereas when others are operative these similarities will be overshadowed by other similarities or differences. If this is right, then in establishing that scientific psychology does (or at least might be able to get away with) recognising, and sometimes appealing to intentional properties that are not locally supervenient, I have established what might be described as a tentative externalist conclusion.

## Chapter 4

# Individualism and Marr's Theory of Vision

### 4.1 Introduction

It is not much of an exaggeration to describe David Marr's theory of vision as contemporary philosophy of mind's favourite psychological theory. Since the publication of his book *Vision*, rarely has a philosopher made a significant claim concerning the nature of scientific psychology or psychological explanation without appealing to Marr's theory to justify her views. In keeping with this fact, much recent discussion of the question of whether or not scientific psychology is individualistic centres on Marr's theory of vision. As might be expected, consideration of Marr's theory has failed to produce any universal agreement. On the one hand there are those writers who believe that Marr individuated the psychological states that figure in his theory non-individualistically, and on the basis of this conclude that scientific psychology - or at least significant parts of it - is, and is legitimately, externalist.<sup>57</sup> On the other hand there are those writers who find no convincing evidence for such a view; for them Marr's theory doesn't present a counterexample to individualism.<sup>58</sup>

Given my concern with the question of whether scientific psychology is individualistic I cannot ignore the discussion of Marr's theory. In this chapter I will reflect on some of the key episodes in this ongoing debate, and will attempt to establish the pro-individualist conclusion that the contents that Marr attributes to the states of the human visual module are locally supervenient. At first sight, this conclusion appears to be at odds with the line of thought that I developed in Chapter 2. However, I believe that a

---

<sup>57</sup> Burge (1986) and Davies (1991), (1992) belong to this category.

<sup>58</sup> Segal (1989), (1991), Egan (1991), (1992), (1994) and Patterson (1996) advance such a view.

reconciliation can be effected. The best way to begin the proceedings is to provide a description of the key features of Marr's theory.<sup>59</sup>

## 4.2 Marr's theory of vision

We have the capacity to discover facts about the world by means of vision; facts such as the shape, size, colour, surface markings and motion of objects in our immediate environment. How do we do this? How is this capacity to be explained? Marr held that underlying this visual capacity is a subpersonal information-processing system housed within the brain. The system in question is the visual module and Marr's theory is an attempt to describe its workings, and in so doing go some way towards explaining our visual capacities. He argues that a complete account of the workings of the visual module will comprise three distinct levels: these being the computational theory, the theory of representation and algorithm and the theory of hardware implementation.

The computational theory concentrates on the semantic and intentional details of the visual module's activity. Its task is to specify the information processing problem solved by the visual module, and the semantic and intentional details as to how it solves this problem. In order to construct an adequate computational theory one must do the following. (i) The information that the visual module takes as input and the information that it generates as output must be indicated. (ii) The visual module doesn't generate output from input in one fell swoop. Rather it makes a series of information processing steps. Consequently, these information processing substeps and the intermediary information that they generate must be described. (iii) The visual module generates information from information by applying mathematical operations or computing mathematical functions, just as I work out the current balance of my bank account from information about its previous state and intervening banking transactions by means of such

---

<sup>59</sup> The most complete expression of Marr's theory can be found in his book *Vision*. Also relevant is his paper 'Artificial Intelligence - a Personal View'. Excellent overviews of subsequent vision research within the broad framework of Marr's approach can be found in Johnson-Laird (1988), Stillings et al. (1987) and Yuille and Ullman (1990).



mathematical operations as addition and subtraction. Consequently, the mathematical operations executed by the visual module (in other words, the mathematical functions that it computes) must be described in precise mathematical terms. (iv) The appropriateness of the information that the visual module generates and the mathematical operations that it employs must be indicated. Such questions as "why is that information and those operations relevant to the task at hand?" and "how do they facilitate the solution of the information processing problem that it is the visual module's function to solve?" must be answered. The appropriateness of the intermediary information generated on the one hand, and the mathematical operations by means of which it is generated on the other, is closely interlinked. What makes a piece of intermediary information relevant is that it is information from which further relevant information can be generated by the mathematical means available to the visual module. And what makes a particular mathematical operation relevant is that it can be employed to generate relevant information from information that the module already has at its disposal. The appropriateness of such information and operations depends crucially on general facts about the world, facts that Marr calls physical constraints. An example of such a physical constraint is that changes in the intensity of light falling on the retina are caused by, and correspond to, such objective features of the external world as boundaries between objects, edges of objects, changes in colour, changes in surface texture, surface contours, and such like. Another example is that physical objects are rigid, not changing their shape and size from one moment to the next. The visual module takes advantage of these constraints with assumptions corresponding to them being hard wired into it. Were these constraints not to hold, the visual module would not be able to generate the information that it in fact generates by means of the mathematical operations that it employs, or what information it did succeed in generating would be of little use or relevance. For example, were it not the case that such features as object edges, boundaries between objects, and the like, caused and corresponded to sudden changes in light intensity falling on the retina, then there would be little point in the visual module's constructing a representation of the location of sudden intensity changes on the

retinal image; such a representation would not contain information from which the shape of the objects impinging on the subjects visual apparatus could be derived. Consequently, the physical constraints that the visual system takes advantage of must be described.

The computational theory tells us nothing about the syntactic details of the visual module's activity. It does not tell us which formal language it employs, or what syntactic rules it applies to the symbols of that language. One and the same computational theory could be implemented in several different ways at the syntactic level, just as the mathematical function of addition could be computed by applying one body of syntactic rules to Arabic numerals, or another body of syntactic rules to binary symbols. The theory of representation and algorithm spells out the syntactic details of the visual module's activity.

Both the very information processing and the very syntactic activity that the visual module engages in could be implemented in many different ways at the physical level. The details as to how that activity is implemented in the brain is not specified by either the computational theory or the theory of representation and algorithm. Rather, the theory of hardware implementation does that.

So we know what the goal of Marr's research is, but what are the details of the theory that he offers? How does he describe the workings of the visual module? Marr's theory is far from complete; in particular, the neural and syntactic details are somewhat limited and sketchy. However, he has much to offer by way of a computational theory. The outlines of that computational theory can be described in the following manner. When we open our eyes to the world, light waves reflected off objects in the immediate environment are focused onto the retina. The intensity of the light falling on the retina will vary from point to point as a result of such factors as differences in the source and strength of illumination across the viewed scene, differences in colour, texture, orientation, and other properties of the various surfaces off which light is reflected, and so on. The retina is a transducer. In response to the light falling upon it the retina generates a two-dimensional array, each value of which represents the intensity of light falling on the corresponding point of the retina. This two-dimensional array is known as the grey coding. The visual module takes as input pairs of

grey codings. As output, the module produces object centred 3-D representations that indicate the shape, size, colour, and motion of whatever objects are in the subject's field of view. This representation can then be employed by other cognitive modules to perform such tasks as object recognition and classification. Hence, the information-processing problem solved by the visual module is that of generating object centred 3-D representations from pairs of grey-codings.

This information processing problem is solved in three distinct stages. In the first stage the primal sketch is constructed. In the second stage the  $2\frac{1}{2}$ -D representation is constructed. And in the third the object centred 3-D representation is constructed. In greater detail, these stages can be described as follows.

1. *The construction of the primal sketch.* As noted above, changes in light-intensity values across the retinal image tend to be caused by, and correspond to, such objective features of the viewed scene as boundaries between objects, object edges, changes in surface texture, changes in colour, surface contours, and the like. Consequently, such changes in intensity-value are potentially a very rich source of the kind of information that the visual module seeks to extract. Hence, it makes sense to represent explicitly the presence and location of significant changes in intensity on the retinal image. This is what the primal sketch does. But how is the primal sketch constructed? To detect intensity changes at a particular scale, a mathematical operator is applied to groups of neighbouring values of the grey codings to generate another two-dimensional array, each value of which represents the gradient of the gradient of intensity at the corresponding point of the retina at the scale in question. Where a positive value in such an array lies next to a negative value, the point between them is known as a zero crossing. Zero crossings correspond to intensity changes. Different sized operators are applied to the grey codings in this manner to detect intensity changes at different scales; small operators will detect small scale local changes, whereas larger ones will reveal the presence of more gradual, larger scale changes. The results of applying these different-sized operators are combined to produce a representation that represents the presence and location of zero crossings that show up at more than one scale. This is the raw primal sketch in which groups of adjacent

zero crossings are represented as edge segments and blobs. Computations are then performed on the raw primal sketch to produce the full primal sketch which represents the global pattern of intensity changes. (For example, locally similar items are clustered to form higher-level units, and boundaries between different regions are detected and explicitly represented).

2. *The construction of the 2 $\frac{1}{2}$ -D sketch.* The second stage of the visual process involves the construction of the 2 $\frac{1}{2}$ -D sketch, a representation that represents the relative distance from the subject of each point on the viewed surfaces together with their orientation relative to the subject. This representation is generated from the primal sketch by means of the execution of several modular processes. These processes include stereopsis and the extraction of depth information from motion information. Both stereopsis and the generation of motion information involve the comparison of distinct representations (primal sketches) and the computation of the displacement of corresponding elements of the representations so compared (where elements of distinct representations correspond with one another if and only if they are caused by and represent the same feature of the objective world). In order for such a comparison process to be performed, a matching process that pairs off corresponding elements of the respective representations must first be executed. This matching process relies upon the three following assumptions. First, that an element of a representation has at most one element of any distinct representation corresponding to it. Second, that elements of distinct representations correspond to one another only if they are similar. And third, that displacements are small and vary smoothly. As a result of making these three assumptions, the matching process pairs an element of a representation with at most one element of any distinct representation, pairs elements of distinct representations only if they are similar, and pairs elements only if they occupy similar positions in their respective representations. That such a matching process is successful, and thus facilitates the subsequent extraction of depth and motion information, is a product of the fact that the assumptions that it relies upon are true, or correspond to real facts about the world.



3. *The construction of the 3-D representation.* In the final stage of the visual process, computational operations are applied to the primal sketch and the 2½-D sketch in order to generate an object-centred 3-D representation of the objects in the viewed scene, a representation that is such as to facilitate object recognition and classification. The 3-D representation represents objects as generalised cones or ensembles of generalised cones, and thus relies on the assumption that our world is populated by objects that have such shapes. The details as to how this final representation is constructed are complicated, but once again the utilisation of motion information plays an important role. Generally speaking, if the motion of two neighbouring features of a primal sketch that are separated by a zero crossing differ over time, then they will lie on different surfaces, for if they lay on the same surface their respective motions would be equivalent. This constraint enables the visual module to determine which zero crossings correspond to object edges (as opposed to, say, changes in colour or surface texture) and thus helps in the construction of the 3-D representation.

That completes my initial account of Marr's theory. Other important aspects of the theory and Marr's general approach will become apparent in due course.

#### 4.3 Burge's argument

We now come to the question of whether Marr's theory is individualistic. Working on the assumption that Marr individuates visual states in terms of their content,<sup>60</sup> this question boils down to that of whether the contents that Marr attributes to the states of the visual module supervene on the subject's intrinsic properties. The central - and agenda setting - argument for a negative answer to this question is due to Burge (1986a).

Burge begins by making the following two claims about Marr's theory:

- (i) 'the theory makes essential reference to the subject's distal stimuli and makes essential assumptions about contingent facts regarding the subject's physical environment' (p. 29).

---

<sup>60</sup> This assumption is widely held. For a dissenting voice see Egan (1991), (1992).



(ii) 'the theory is set up to explain the reliability of a great variety of processes and sub-processes for acquiring information, at least to the extent that they are reliable' (p. 29). In other words Marr's theory is not "success neutral".

He takes (ii) to imply a third point, namely that:

(iii) 'the information carried by representations - their intentional content - is individuated in terms of the specific distal causal antecedents in the physical world that the information is about and that the representations normally apply to' (p. 32).

My description of Marr's theory should make it clear that claims (i) and (ii) are both true. However, it is less than clear that (ii) implies (iii). There are two routes of escape that the individualist might attempt to follow. The first involves questioning Burge's understanding of the nature of the success-orientation of Marr's theory. The thought is that although Marr takes the human visual system to be reliable, he is not committed to viewing as reliable the visual systems of all our possible twins.<sup>61</sup> The second option is to adopt what Davies (1991) calls the revisionary strategy. This involves ascribing to visual representations non-workaday contents that would be veridical across a wide range of environments, despite the fact that their typical causal antecedents would vary from one environment to the next. The thought is that the contents of my visual representations are such that there are possible environments quite unlike our world where, were I placed in them, my visual system would be reliable even if its representations had the contents

---

<sup>61</sup> Patterson (1996) follows this line. She argues that for Marr it would be an open empirical question whether the visual systems of our possible twins were reliable. She appears to lean towards the view that in those environments where my twin's visual representations would have to have contents that diverge from their counterpart representations in me for his visual system to be veridical, his behaviour should (and probably would) lead Marr to conclude that he was systematically misrepresenting his world (and thus that his representations had the same content as mine).

they actually have. Hence, one can hold onto the assumption of success without going externalist.<sup>62</sup>

I will argue that neither claim (i) nor claim (ii) (or the conjunction of them) supports an externalist conclusion. As my argument draws upon points that emerge most clearly in the context of a discussion of some putative counterexamples to the individualist thesis I shall not present this argument just yet. Suffice it to say that, as things stand, it is far from clear that Burge has a compelling case. However, Burge offers much more by way of argument. For example, he presents the following chain of reasoning<sup>63</sup>:

(1) The theory is intentional. (2) The intentional primitives of the theory and the information that they carry are individuated by reference to the contingently-existing physical items or conditions by which they are normally caused, and to which they normally apply. (3) So if these physical conditions and, possibly, attendant physical laws, were regularly different, the information conveyed to the subject and the intentional content of his or her visual representations would be different. (4) It is not incoherent to conceive of relevantly different (say, optical) laws regularly causing the same non-intentionally, individualistically individuated physical regularities in the subject's eyes and nervous system. It is enough if the differences are small; they need not be wholesale. (5) In such a case (by (3)) the individual's visual representations would carry different information and have different representational content, though the person's whole non-intentional physical history (at least up to a certain

---

<sup>62</sup> Segal (1989) adopts this strategy when, in connection with Burge's crack-shadow example (see below) he attributes to both the actual and counterfactual P's representation O the content *crackdow*, this being a content that applies equally to both cracks and shadows.

<sup>63</sup> In this argument, the controversial claim (iii) appears as premise (2) without any additional support. However, this premise does appear in a dialectical context where it is easier to appreciate its power and plausibility, what would be involved in rejecting it, and just how difficult it is to construct a convincing defence of an individualistic reading of Marr. Moreover, the argument presents the individualist with something more tangible to grapple with. Hence there is profit to be had from focussing one's attention on this argument.

time) might remain the same. (6) Assuming that some perceptual states are identified in the theory in terms of their informational or intentional content, it follows that individualism is not true for the theory of vision.

What are we to make of this argument? What is needed to make it convincing is a plausible example. That is, what the externalist needs to do is present us with a sufficiently worked out example of a humanoid individual with a visual system that is individualistically and non-intentionally just like ours, who lives in, and is well adapted to, an environment that is unlike ours to such an extent that it would be indefensible to ascribe to her visual representations just the same content that Marr would ascribe to the corresponding representations in us. The problem is that it is difficult to come up with a plausible such example. In the search for such an example the externalist is pulled in opposite directions. On the one hand, the imagined scenario needs to be plausible. This requirement pulls the externalist towards describing a case where the imagined twin's environment is only subtly different from ours. The more outlandish the example the less plausible the claim is that a creature with a humanoid form could be well adapted to, and prosper in, the imagined environment.<sup>64</sup> On the other hand, the closer to our world the imagined environment is, the less pressing will be the need to ascribe to the visual states of its inhabitant's contents that diverge from the contents of the corresponding states in us in order to account for their capacity to find out about the nature of their world by means of vision. However, all this notwithstanding, there is an example worthy of examination. This is a modified version of the famous shadow-crack case that is presented by Burge as an independent refutation of individualism. I will examine this case and argue for the conclusion that it can be successfully accommodated by the individualist.

Before examining the shadow-crack case, I will make a brief digression. According to the individualist, the contents that Marr

---

<sup>64</sup> It is important that the creature is well adapted to its home environment, otherwise there will be strong grounds for concluding that its visual states systematically misrepresent their distal causes (perhaps by having just the same contents as their analogues in us).

attributes to our visual states supervene on our intrinsic physical properties. This rules out the possibility of my having a twin the content of whose visual states diverges from that of their counterparts in me. Thus, one way of defeating individualism is to present a compelling counterexample. However, the externalist is not committed to the idea that there are any such counterexamples; it is open for her to believe that any environment sufficiently different from ours to generate alternative visual contents would be such that none of our twins could be well adapted to it. Such an externalist will concede that all well adapted twins will share the same visual contents. However, (and this is what makes her an externalist) she will hold that this content-equivalence is partly the product of similarities in the twins' respective home environments and their relations to it, and not wholly the product of their internal physical similarities. But there is a major problem for the externalist who adopts this position and thus abandons, or condemns as fruitless, the search for twins with divergent visual contents. For, short of producing a twin case analogous to the folk psychological examples of Putnam and Burge, it is difficult to see how one could vindicate the thesis that Marr individuates non-individualistically. To see this consider the following.

As we have seen, Marr was concerned with the workings of the human visual system and did not have in mind any creatures living in a world other than ours. He portrays the visual system as generating representations that represent the extra-cranial world as being a certain way; in other words, the visual system sticks its neck out as to how the extra-cranial world is. Marr held that the visual system is by and large successful in representing the world as it really is; for him its representations are largely veridical. Thus he is committed to the existence of a good deal of causal covariation between properties instantiated in the external world and visual representations that represent those properties as being instantiated in the external world. In other words, for Marr, visual representations that represent their distal cause as having the property P tend to covary with instances of that property. An upshot of all this is that Marr describes the contents of visual representations in terms that refer to properties instantiated in the extra cranial world. But from this alone no externalist conclusion can be derived.



After all, Descartes, that paradigmatic individualist, described the contents of his thoughts in terms that referred to properties that, if instantiated at all, are instantiated in the extra mental world. And after becoming satisfied that God exists and has a benevolent nature, Descartes came to the view that there is a good deal of causal covariation between his perceptual states and the external properties that they represent as being instantiated in the world. What made Descartes an individualist is that he held that the content of his thoughts is determined solely by the intrinsic nature of his mind.

What would indicate that Marr was committed to externalism would be the presence in his work of comments that committed him to some externalist theory of content (perhaps a causal covariation theory or a teleological theory of the sort advanced by Millikan (1984)). But there is no reason to expect Marr, or any other scientific psychologist, to make such comments or reflect upon the nature and origins of content; it is philosophers, rather than psychologists and cognitive scientists, who tend to do that sort of thing. There would be a greater likelihood of Marr saying something explicit on these questions if he were confronted with twins who inhabited different environments and had to reach some decision as to what contents to attribute to their visual representations. If Marr had been faced with the need to make such decisions, he may well have engaged in some general reflections on the nature of content, either to settle the question of what contents to attribute to our twins, or to justify a decision that he had made. But of course, Marr didn't have to worry about humanoid subjects living in, and being well adapted to, environments other than our own. But if Marr, and psychologists in general, don't - and don't need to - address themselves to questions as to the nature and origins of content, the only way to establish their externalism would be to produce an example of twins to which they would be unlikely to attribute the same contents because to do so would be implicitly to commit themselves to some bizarre and indefensible theory as to the nature and origins of content. The analogy here is with folk psychology. Ordinary folk are forever attributing contents to the mental states of themselves and their fellows in the course of describing, explaining, predicting, and making sense of their behaviour and mental life. However, ordinary folk do not explicitly commit themselves to any theory as to the



nature and origins of content, *a fortiori* they do not explicitly commit themselves to any externalist theory of the nature and origins of content. This fact doesn't stop folk psychology from being, as it were, implicitly externalist. Folk psychology is implicitly externalist in the respect that its practitioners attribute to our mental states contents that it would be indefensible to attribute to our twins.<sup>65</sup> Now Marr's theory could be implicitly externalist; that is, it might attribute to our visual states contents that it would be indefensible to attribute to all of our metaphysically possible twins. In order to establish that Marr's theory was externalist in this way, what is needed is a plausible example of a humanoid creature living in, and well adapted to, an environment different from ours to such an extent that it would be indefensible to attribute to his visual representations just the same contents that Marr would attribute to the corresponding states in its human twin. Hence, there is good reason for those with externalist leanings to launch themselves into a search for a plausible example of twins with divergent visual contents. However, there are risks involved for, quite apart from the prospect of returning empty handed, a consideration of putative counterexamples to individualism might only serve to emphasise features of Marr's theory and approach that ultimately tell against externalism.

So the question arises: are there any plausible examples? A case worthy of consideration is a modified version of Burge's shadow-crack example.<sup>66</sup> This runs as follows. An individual P lives in an environment in which there are both small shadows and similar-

---

<sup>65</sup> Indefensible for the following kind of reasons: because it would commit one to some bizarre theory as to how our twins' states got the content that our states have (how could Oscar<sub>2</sub>'s thoughts be water thoughts when he lives in a water-free world?); or because it would represent our twins as systematically misrepresenting their world when there is no more of a ground for the conclusion that they misrepresent their world than that we misrepresent ours; or because it undermines the attributions made to us. (If Oscar's thoughts can be water thoughts what is stopping our thoughts from being twater thoughts? Surely not that there is no twater around here. For if that is a good reason for regarding our thoughts as water thoughts then, by parity of reasoning, there are good grounds for attributing twater thoughts to our twins on Twin Earth).

<sup>66</sup> This appears both in Burge's 'Individualism and Psychology' and in his 'Cartesian Error and the Objectivity of Perception'.

sized and shaped cracks. Tokenings of a representation of type T in P are typically caused by shadows in normal conditions. Marr would say of tokenings of T that they represent their distal cause as a shadow; in other words T has the content *shadow*. T's are occasionally caused by cracks and in such cases a crack is misrepresented as a shadow. When a crack causes a tokening of T, P invariably has no dispositions to distinguish the instance of a crack from instances of shadows. In effect, P's visual system cannot distinguish between shadows and cracks in normal circumstances. However, it can distinguish between shadows and cracks in ideal conditions. Next consider P in a counterfactual situation where he is individualistically and non-intentionally just as he is in the actual situation. P's counterfactual environment is somewhat different from his actual one, for in it there are no shadows, and tokenings of T are standardly caused by cracks. If P were to be confronted with a shadow in these circumstances (perhaps counter-normically) it would be either invisible to his visual system or would produce a tokening of some representation other than T. Marr would, or should, attribute to T in the counterfactual case the content *crack*. It would be absurd to ascribe to it the content *shadow* given the facts about its standard etiology.<sup>67</sup> Thus we have a case of twins who, from the point of view of Marr's theory, are in different visual states due to differences between their respective environments.

I shall now attempt to establish that this is not a convincing argument; it does not establish that Marr's theory is externalist. It is

---

<sup>67</sup> To repeat, this is a modified version of the example presented by Burge. His concern in presenting this example is to produce a general argument against individualism as a theory about visual states. He says of his argument that it 'is independent of the theory of vision that we have been discussing [that is, Marr's theory].. It supports and is further supported by that theory' (1986a, p. 43). Thus he is not specifically talking about Marr's theory, nor are the representations or states that he describes necessarily those (or just like those) that figure in Marr's theory. For example, it may well be the case that the states or representations that Burge describes are intended to be understood as personal level states rather than sub-personal states (unlike the states and representations that feature in Marr's theory). Moreover, Burge initially presents an abstract version of the case which talks of the properties O and C. Cracks and shadows appear when he attempts to provide a concrete instantiation of his anti-individualist argument.

quite plausible to say that Marr would ascribe the same content to T in the counterfactual case as he would in the actual one, for the contents that he attributes in accounting for our visual capacities need not be as specific as the argument supposes. As a first step in establishing this conclusion consider the case of Putnam's twins once more. Both Oscar and Oscar2 sometimes token a visual representation of intrinsic type R. In Oscar, tokenings of R are caused by water, and in Oscar2 they are caused by twater. Thus the types of stuff that R is typically caused by and applied to differs from one twin to the next. Yet this fact alone doesn't imply that R's in Oscar mean (from the point of view of Marr's theory) one thing, and R's in Oscar2 something else. Clearly the contents that Marr attributes to visual representations are not so specific as to distinguish between water and twater. Thus Marr's theory is such that not every difference in typical distal cause counts for a difference in content. This should make us alive to the possibility that the difference in typical distal cause between the actual P's T's and the counterfactual P's T's doesn't count for a difference in content.

It might be argued that the crack-shadow case is fundamentally different from the water-twater case for the following kind of reasons. Firstly, the range of properties that the visual system is sensitive to, and represents as such, is rather limited, and does not include such properties as that of being water and that of being twater. The properties the visual system is sensitive to and does represent as such are shape, size, colour, texture properties, and the like. In terms of such observable properties, water and twater are alike, thus, from the Marrian perspective, the typical distal cause of R in Oscar is just the same as the typical distal cause of R in Oscar2. Secondly, cracks and shadows do differ from the point of view of Marr's theory as can be seen by recalling what Marr regards as being one of the fundamental tasks of the visual system. Changes in light intensity values in the retinal image are caused by a number of distinct physical phenomena including object edges, changes in surface orientation, changes in texture, shadows and so on. One of the major tasks of the visual system is to disentangle these various physical causes, determining which features of the retinal image (and primal sketch) are caused by object edges, which by shadows, and so on. Were the visual system not capable of successfully

accomplishing this task, it could not produce the veridical output representations that it in fact produces. A consequence of this is that the visual system cares about the difference between cracks and shadows and must be able generally to see shadows as shadows (and cracks as cracks) whenever it comes across them.

The above points generate the thought that contents such as *crack* and *shadow* are precisely the kind of contents that Marr would attribute to visual representations (unlike *water* and *twater*), and that for their respective visual systems to be successful the actual P's T's would have to represent their distal cause as a shadow whereas the counterfactual P's T's would have to represent their distal cause as a crack.

The above line of argument is unconvincing. It is consistent with the success of both the actual and the counterfactual P's visual system and the fact that for Marr the visual system disentangles the effects of - and thus distinguishes between - such disparate physical phenomena as object edges and shadows, that T's have identical contents in both the actual and the counterfactual case. What the shadows and the cracks that Burge describes have in common is that they are both thin, dark, surface marks. It is my contention that Marr would attribute to both the actual and the counterfactual P's T's some content such as *thin, dark, surface mark*. That is to say that whenever actual or counterfactual P's visual system generates a T it represents the surface before it as having a thin dark mark on it.<sup>68</sup> Here's why. Just because the visual system must generally be able to distinguish between shadows and other types of physical phenomena, and thus must be able to represent shadows as such, it doesn't follow that there are not some kinds of shadow that it doesn't represent as shadows, or that it is never neutral as to whether the distal cause of a surface feature it has detected is a shadow or some other physical phenomenon. Sometimes a visual system will detect a surface feature - typically small and insignificant - that it is unable to determine much of the nature of. In such a case it is useful to have a representation that specifies as much as is known about the surface

---

<sup>68</sup> In attributing such a content to T, I am not adopting what Davies (1991) describes as the revisionary individualist line, as I am not attributing a non-workaday disjunctive content to T. In particular, I am not (unlike Segal, 1989) attributing the disjunctive content *crackdow* (i.e. *shadow-or-crack*) to T.



feature in question but is noncommittal or neutral as to such details as whether it's a shadow, an object resting on the surface, a crack, or whatever. If the visual system "stuck its neck out" in such cases it would be liable to error or, at the very least, would be making an arbitrary, unjustified decision as to what was before the subject. The kind of case I am imagining is the analogue of the familiar personal level case where we see a viewed surface as having a small dark mark on it but we can't make out what the mark is; that is we can't make out whether it is a shadow, a crack, another object resting on the surface (e.g. an insect), a change in surface texture, or whatever. In such circumstances we just see the surface feature as a small, dark mark. To jump to the conclusion that the surface feature was, say, a shadow would be premature and unjustified, making one prone to error and misrepresentation.

My point is that in the crack-shadow case Marr would - or at least could have - regarded the T's as having the content *thin, dark surface mark*. Just because the actual P's T's tend to be caused by shadows, it doesn't automatically follow that Marr would regard T as having the content *shadow*. It would be beneficial to a visual system to have the resources to register the presence of surface features without making any significant conclusions as to their nature (beyond such properties as their shape and size). That we have such a case of noncommittal representation in the Burge example is suggested by several features of that example, namely that the shadows and cracks are small (it is difficult for the visual system to determine the nature of small surface features, especially from a distance), and that the subject isn't very interested in the surface feature (as is indicated by the fact that the subject has no dispositions to employ any other sensory modality or move closer so as to rule out the possibility of error or find out more about the surface feature). In short, in portraying the shadows and cracks as minor surface features that the subject isn't very interested in, Burge removes any motivation for regarding the actual P's T's as representing its distal cause as a shadow and the counterfactual P's T's as representing its distal cause as a crack. Moreover, the familiar visual experience of seeing a minor surface feature just as a small, dark mark (rather than as a token of a specific type of physical phenomenon) suggests that the visual system is



often fairly neutral as to the nature of the distal stimuli impinging upon it.

Another relevant point is that, so to speak, relatively neutral representations do figure prominently in Marr's theory. For example, the edge segments that are amongst the primitives that make up the primal sketch are caused by, and correspond to, a whole range of physical phenomena (such as object edges, changes in surface orientation, and the like), but whenever such a primitive is tokened it is neutral as to the physical nature of its distal cause. In so far as edge primitives say anything about their distal cause it is that there is a significant physical feature out there the nature of which hasn't yet been determined. (Segal, 1989).<sup>69</sup>

In conclusion, then, Burge hasn't presented us with a convincing anti-individualist example, for he has given us no reason for thinking that Marr would attribute the content *shadow* to the actual P's T's in preference to the less specific *thin, dark surface marking*. The latter attribution would appear to be intuitively the most appealing given the subject's lack of interest in the specific nature of the surface feature and the insignificance of that nature to him, along with the familiar fact that we often see surface features just as thin dark surface markings.<sup>70</sup> And if the actual P's T's mean *thin*,

---

<sup>69</sup> It might be argued that there is a key difference between edge segments and the representations Burge imagines, for the latter, unlike the former, tend to covary with a single type of physical phenomena. In response, this objection misses the point of my appeal to the example of edge segments. The point of this appeal is only to indicate that Marr presents the visual system as having the capacity to be relatively unspecific in how it represents the external world to be when it is not in the position to determine all that much about the specific nature of its distal stimuli. Moreover, the content that I attribute to the T's doesn't violate any covariation requirement.

<sup>70</sup> That the actual P's T's are typically caused by shadows gives no grounds for attributing a more specific content than *thin dark surface marking*. Consider the case of personal level visual experience where the surface features that I see as thin dark surface markings are sometimes one kind of physical phenomena and sometimes some other kind. Had I lived in a world where the distal causes of those visual experiences were always one single type of physical phenomena (say small cracks) the content of my visual experiences would not thereby be different from what they

*dark surface marking* then, by parity of reasoning, so do the counterfactual P's T's.

It might be objected that Davies (1991) provides a refutation of my conclusion. In his assault on revisionary individualism, Davies argues that that brand of individualism cannot cope with a significant feature of the Burge example, namely 'that in ideal circumstances P is quite able to discriminate a C (a crack) from an O (a shadow); it is simply that P rarely sees Cs (cracks), and when he does it is in non-ideal circumstances' (p. 475). Davies argues that the only way of accounting for this discriminatory capacity is to attribute to T the content *shadow*. How should we respond to this?

A first point is that the content attribution that I am arguing for is to be distinguished from the disjunctive content that the revisionary individualist argues for. Hence the failure of the latter doesn't imply that my position is not tenable. So can the attribution of the content *thin, dark surface marking* make sense of P's discriminative abilities? Presumably what is ideal about the ideal circumstances is that in them P can reliably discriminate between shadows and cracks. Having this ability does not imply that P sees the shadows as shadows and the cracks as cracks, but only that how he sees the latter is different from how he sees the former. These circumstances are not ideal as such but only ideal relative to the task of discriminating shadows from cracks. Thus the normal circumstances in which P lives out most of his shadow-and-crack viewing life are not non-ideal as such, but only non-ideal relative to the task of telling shadows apart from cracks. In other words, they are not the analogue of pitch blackness for us, that is, a type of circumstance in which our visual system cannot, and has not evolved to, work in. There is no evidence that Burge conceives the circumstances in which P's visual system responds to both shadows and cracks by producing T's as being abnormal or as constituting circumstances that that visual system has not evolved to cope with. Thus we can say that in normal circumstances P cannot discriminate between shadows and cracks, a fact that is not without significance, for if P had such a discriminative capacity it would be implausible to claim that T has a content (*thin, dark, surface marking*, for example) that applied to both shadows and

---

in fact are. Why would it be any different in the case of sub-personal visual representations?

cracks. So the question is: is P's capacity to discriminate between shadows and cracks in ideal circumstances such as to imply that T does not mean *thin, dark surface marking*? Burge does not tell us much about what happens in the ideal circumstances when P is confronted with a shadow (or a crack). There are two possibilities consistent with what he says. First, it is consistent with what he says that in ideal circumstances neither shadows nor cracks cause T's, but rather that shadows reliably cause tokenings of a second representational type T'', and that cracks reliably cause tokenings of a third representational type T'''. In such a case the attribution of the content *thin, dark surface marking* to T does not make P's capacity to discriminate between shadows and cracks in ideal circumstances unintelligible. On the contrary, that capacity can be accounted for by attributing a second content to T'' and a third content to T'''. Indeed, if facts are as described it would be counterintuitive to ascribe to T the content *shadow*.

That this first scenario is not at all outlandish is indicated by a familiar everyday analogue. In the normal course of things, we view surfaces from a distance, and as a consequence we are often not able to determine much as to the nature of small surface features. Often, when we see small surface features, we cannot make out whether they are cracks, shadows, or whatever; typically what we see such surface features as is small dark surface markings (or something along those lines). If we are disposed to find out more as to the nature of such a surface feature we will move closer so as to discover more as to its nature. When we move closer we typically come to see the surface feature differently; we might come to see it as having a different (or perhaps a more specific) shape or colour, as having - or not having - depth, as having a specific texture, and so on. In such cases, how we come to see the surface feature will depend a lot on its intrinsic nature; thus when we move closer to cracks we typically see them differently from how we see shadows when we move closer to them, despite the fact that how we see the latter from a distance is just how we see the former from a distance. In this kind of familiar case, what goes on when we are in ideal circumstances - that is, circumstances ideal relative to the task of discriminating shadows from cracks - in no way tells against the idea that in normal

circumstances we see both the shadows and cracks in a neutral and non-specific way.

The upshot of the above reflections is that the content that I have attributed to T is consistent with P's having a capacity to discriminate between shadows and cracks in ideal circumstances. However, the specifics of the Burge case might take a second form. In such a case, shadows reliably cause T's in both the normal and the ideal circumstances, whereas cracks cause T's only in the normal circumstances, causing some other representation in ideal circumstances. It might be thought that this scenario would tell against the content attribution that I have been arguing for by suggesting that T has the content *shadow* and the representation caused by cracks in ideal circumstances the content *crack*. But this is not obviously the case. In normal circumstances P's visual system could represent both shadows and cracks as thin dark surface markings, that is, represent both types of phenomena in neutral and non-specific terms; yet on P's moving closer it could uncover more specific facts about the nature of what was before P when what was before P was a crack but not when it was a shadow. For example, on P's moving closer his visual system might be able to determine more about the specific shape or colouration of cracks - something that it can't do with respect to shadows. An important point to note is that P's visual system might be uncovering properties of cracks that shadows do in fact have, or could have. Thus there is no reason to conceive of the representations caused by cracks as having the content *crack*. An upshot of this is that once again we can account for P's ability to discriminate between shadows and cracks in ideal circumstances without attributing to T the content *shadow*, a content that it might be difficult to justify ascribing to the counterfactual P's T's given the fact that they are never caused by shadows.

In short, then, P's capacity to discriminate between shadows and cracks in ideal circumstances does not in and of itself tell against the attribution of the content *thin dark surface marking* to T. Therefore such an attribution is consistent with the shadow-crack example as described by Burge. Of course this doesn't rule out the possibility that Burge's scenario could be developed in such a way as to cause problems for the individualist, but that development has not been executed by Burge. I suspect that any such development would have



to make the difference between P's actual and counterfactual environment much more graphic and wholesale than that implied by Burge (recall that Burge presents the case as being one where the difference between the respective environments is only minor). However, as I stated above, the cost of making such a move is a reduction in plausibility. Rather than attempting to develop the shadow-crack example I shall turn my attention to a case of twins who inhabit environments that differ in a graphic and wholesale way.

#### 4.4 Circles and squares: a more graphic putative counter-example

Suppose that S1 is a state of the human visual module that is normally caused by square-shaped objects and rarely by objects of any other shape. In addition, suppose that whenever a subject tokens S1 she subsequently has a visual experience *as of* a square, an experience with the content *square*. Whenever a subject takes such an experience at face value she subsequently behaves in a way that would be appropriate were there a square-shaped object before her, and inappropriate if otherwise. What is the content of S1? The obvious answer is *square* (or *square-shaped object*). Now consider S2, another state of the visual module. S2 tends to covary with circular-shaped objects and causes visual experiences with the content *circle* (experiences that, when taken at face value, manifest themselves in behaviour appropriate to circular-shaped objects). What is the content of S2? The obvious answer is *circle* (or *circular-shaped object*).

In a distant world inhabited by creatures intrinsically very much like us, optical laws and other external conditions are such that it is circular-shaped objects that typically cause tokenings of S1 in the locals, and square-shaped objects that typically cause tokenings of S2. What is the content of S1 and S2 in the inhabitants of this world? One (externalist) answer is that S1 has the content *circle* (or *circular-shaped object*) whereas S2 has the content *square* (or *square-shaped object*). That answer suggests the following counterexample to individualism. Edgar is a normal Earthly subject who, confronted with a square-shaped object, tokens S1 in response. Edgar has a twin who is an inhabitant of the above described distant world who is



currently confronted with a circular shaped object (an object that has caused him to token S1). The thought is that Edgar's S1 has the content *square*, whereas the corresponding state in his twin has a quite different content (namely *circle*). Thus, the content of the states of the visual module is not locally supervenient.<sup>71</sup>

Is this a convincing counterexample to the individualist's thesis? I will argue for a negative conclusion; Marr would (or should) attribute the content *square* to S1 in both us and our other worldly counterparts, which would thus imply that the latter systematically misrepresent the shape of objects in their world. There are two basic arguments for this conclusion, one of which appeals to visual experience, and the other of which appeals to behaviour.

Here is how the argument from visual experience goes. The visual module plays a prominent role in the etiology of visual experience, so that how the world is presented to me in experience (the content of my experiences) will depend upon how my visual module represents the world to be. If I have a visual experience *as of* a square, then that experience will be the product of my visual module's representing its distal stimuli as being a square-shaped object. *Ceteris paribus*, had my visual module represented the distal stimuli as having some other shape, then I would not have had the visual experience that I in fact had. In general, then, there is a systematic and coherent relationship between the output of the visual module and the content of visual experience. Consequently, part of Marr's task is to explain how the visual module manages to detect and represent those features of the world that our visual experiences represent. Marr will fail in this task if the contents that he attributes to the states (particularly the output states) of the visual module aren't appropriately related to the contents of visual experience, for he will have left it a mystery how the visual module enables us to see the world as we see it.

The role that the visual module plays in the etiology of our visual experiences (or, alternatively, in enabling us to see the world as we see it) implies that if the contents of the respective visual experiences of twins are identical then so will be the contents of the states of their

---

<sup>71</sup> This twin scenario is closely based upon one developed by McGinn (1989) in his discussion of the question of whether the content of visual experience is locally supervenient.

respective visual modules, for any divergence in the contents of the latter states would show up in visual experience. In connection with our specific example, if Twin-Edgar's experience is *as of* a square, then the state S1 in him must have the content *square* despite the fact that that state covaries with circularshaped objects. To deny this is to make it a mystery how Twin-Edgar came to see the world as he saw it; if S1 had the content *circle* then surely he would have had a visual experience *as of* a circle.

As yet, none of this causes the externalist any problems, for it hasn't been established that the visual experiences of twins have identical contents. However, the assertion that the content of visual experience is locally supervenient has considerable intuitive power. Visual experiences are states of consciousness; they have phenomenological or qualitative properties. There is something *that it is like* to have an experience *as of* a square. Now what seems undeniable is that the qualitative character of an individual's psychological states is locally supervenient; the psychological states of twins are qualitatively identical no matter how much their home environments diverge. In other words, what it is like to be you is determined by how things are within your skin. Therefore, the qualitative character of Edgar's visual experience when he sees the square shaped object will be the same of that of his twin when he is confronted by the circular-shaped object. Consequently, if the content of a visual experience supervenes upon its qualitative character, then, given the transitivity of the supervenience relationship, the externalist's case collapses. And it has to be noted that it is intuitively highly plausible to claim that if someone has a visual experience with the same qualitative character as my experience *as of* a square, then that experience must also be *as of* a square regardless of the nature of the home environment of the subject in question. Quite generally, it would take some argument to unseat the intuition that the content of an individual's visual experiences supervenes upon the qualitative character of those experiences and thus upon her intrinsic physical properties. I am not going to get embroiled in the question of whether such a supervenience relationship holds<sup>72</sup> but

---

<sup>72</sup> McGinn (1989) argues for the thesis that the content of visual experience supervenes upon qualitative character. Davies (1992) disagrees. For him, the supervenience relationship is of the within world-within species variety.

merely note that visual experience is a potential source of trouble for the claim that the square-circle case constitutes a counterexample to individualism. To recapitulate: if the content of visual experience is locally supervenient then Edgar's visual experience will have just the same content as that of his twin. And if their visual experiences have just the same content then so will their respective tokenings of S1, given the role that the visual module plays in enabling us to see the world.

My central argument for the conclusion that the square-circle case does not constitute a convincing counterexample to the individualist's thesis involves an appeal to behaviour. The basic idea is that the behaviour of Twin-Edgar towards circular-shaped objects (being just the same as that of his Earthly twin towards square shaped objects) suggests that he systematically misrepresents them as squares. Hence, S1 has the content *square* in both twins, environmental differences notwithstanding. In arguing thus I follow in the footsteps of both McGinn (1989) and Segal (1991).

Visual systems are not merely onboard entertainment systems. They perform a very important function, namely that of providing the subject with information about the nature of her world, information that it is useful for her to have if she is going to satisfy her needs and desires and generally prosper in the world. In order to perform this function it is important that visual systems are reliable; their pronouncements as to the nature of the external world must generally be veridical. But it is equally important that the subject is able to act on the basis of how she sees the world. In general, how we see the world shows up in our behaviour. Moreover, it is typically the case that how we behave on the basis of our visual states is appropriate given the content of those states. Thus, if I see an object before me as having a square shape, I will behave towards it in a manner that would be appropriate or sensible were it in fact square-shaped, but not if it were any other shape. In behaving thus, my behaviour is coherently related to the contents of my visual states. Consequently we can, and often do, read off facts about the contents of the visual states of our fellows from facts about their behaviour. If a creature couldn't behave in a manner that was coherently related to the contents of its visual states then there would be little point in its having a visual system; its visual system would play no role in

enabling it to prosper in the world. And if a creature could, but generally didn't, behave in a way that was coherently related to the contents of its visual states, then it might as well not have a visual system.<sup>73</sup>

Now consider Twin-Edgar once more. He will behave in response to circular-shaped objects just as Edgar does in response to squares. For example, in describing the shape of the object before him when he is confronted with a circular-shaped object, he will do such things as trace a square in the air with his hand, or draw a square on a piece of paper. And he will attempt to circumnavigate circular shaped objects by following a square-wise path (a form of behaviour that will either result in him colliding with the object in question or expending excessive energy by going further out of his way than he need have done). In general, behaviour that state S1 is implicated in the etiology of will be appropriate to squares (but not to circles), and that behaviour will be coherently related to the content of S1 if that state has the content *square* (but not if it has the content *circle*).

The upshot of all this is that if Twin-Edgar's behaviour is coherently related to the content of his visual states, then S1 will have the content *square*, and he will be systematically misrepresenting the shape of circular-shaped objects. Thus the externalist has to argue that this is a case where there is a breakdown in the relationship between visual content and behaviour. Here, behaviour pulls us in one direction, and etiology in the other, so that we must choose between the two (or, more accurately, determine which way Marr would jump).

In the case of Earthly humans, the subjects of Marr's reflections, there is no such conflict between etiology and behaviour; Marr did not have to make any decisions on cases anything like the one that

---

<sup>73</sup> Of course we don't always behave in a manner that is coherently related to the content of our visual states. Sometimes intervening factors prevent us from behaving as we normally would, as when we lose control of our limbs through disease or injury. Sometimes we do not take our visual experiences at face value. And sometimes we behave in such a way as to conceal how we see the world from our fellows. Yet such phenomena are exceptions to a general rule. Indeed, in the latter two kinds of case, the subject's behaviour will be coherently related to the content of her visual states along with that of the beliefs and desires that are implicated in her behaving the way she does.



concerns us now. Thus we will find no precedents in his work. However, I think that there are good reasons for thinking that his approach is such that he would have sided with behaviour in this kind of case, and thus that he would have attributed the content *square* to Twin-Edgar's state S1. Such an attribution has the advantage of cohering with commonsense, for I take it that most ordinary folk would accept McGinn's claim that "aftermath trumps etiology", and would thus view Twin-Edgar as being something of a Mr. Magoo.

In determining how the visual module works, Marr relies heavily on the behaviour of human subjects; for example, their behaviour in normal perceptual situations, their behaviour in experimental settings, their behaviour in laboratory conditions where normal physical constraints have been violated, and so on. In so relying on behaviour, he operates on the assumption that behaviour can tell us an awful lot about the workings of the visual system, for example, what information processing problems it solves, what physical constraints it takes advantage of/what assumptions it employs, what features of the world it is sensitive to and represents, and so on. This implies that he assumes that behaviour can tell us an awful lot about the contents of the states of the visual module, for to represent that module as solving a particular information-processing problem, or as relying on a particular assumption, is to make a claim about the contents of the representations that it generates. Thus Marr assumes that there is a systematic and coherent relationship between the goings on in our visual module and the manner in which we behave, and, moreover, that there is such a relationship between the contents of our visual states and the behaviour that they subsequently cause. Indeed, it is difficult to see how Marr could proceed in any other way, for if he were to abandon his reliance on behaviour he would have no hope of constructing a viable theory of vision. It's not as if he can cut open our heads and see what is going on inside.

What Marr's reliance on behaviour (and his assumption that there exists a coherent relation between the contents of our visual states and our behaviour) suggests is the following. Since Twin-Edgar behaves in just the same way as his Earth bound twin, Marr would attribute just the same contents to his visual states. In particular,



Marr would take his behaviour vis-a-vis circular-shaped objects as evidence that he systematically misrepresented them as square-shaped; to attribute any other content to state S1 would be to ignore all the behavioural evidence.

But how does this cohere with the success-orientation of Marr's theory? It might be argued that in addition to his assumption about behaviour, Marr assumes that the visual module generally produces veridical representations. In the case of the Earth bound Edgar, these two assumptions can jointly hold true but not in the case of Twin-Edgar (and his fellows). So, in connection with Twin-Edgar, Marr has to abandon one of his assumptions, and who is to say that it would not be the one that there is a systematic and coherent relationship between the contents of his visual states and the behaviour that they issue in? Quite apart from the point that Marr can't afford to abandon his assumption about behaviour if he is to have any hope of uncovering the workings of Twin-Edgar's visual system, there is an answer to this challenge. Although Marr takes our visual module to be largely successful in correctly representing the world (and sets out to explain this success), it is no part of his position that visual modules must be, or are inevitably, successful in this respect. He is free to take it as an open empirical question just how successful a creature's visual system is, a question that can only be settled by observing the behaviour of the creature in question (Patterson, 1996). The behaviour of us Earth bound humans indicates that our visual states are largely veridical. But matters are somewhat different when it comes to Twin-Edgar and his fellows. With respect to circular-shaped objects, they will systematically engage in a pattern of behaviour that is wholly inappropriate, resulting in their crashing into such objects, their dropping them, and general chaos. Heaven knows what calamities will ensue if Twin-Edgar attempts to go crown green bowling.<sup>74</sup> To Marr all this would surely indicate that

---

<sup>74</sup> The fact that Twin-Edgar's interactions with circular-shaped objects result in such disastrous consequences serves to block a certain line of response that the externalist might try to develop. The argument I have in mind runs thus. Because of the nature of his home environment, we shouldn't automatically assume that when Twin-Edgar traces a square with his hand (or draws a square) his behaviour has the same significance as it would have here on Earth. Perhaps Twin-Edgar's square tracing behaviour means *circle*, thus implying that S1 could have the content *circle* without

Twin-Edgar and his fellows misrepresent the shape of circular-shaped objects, and that their visual modules are far less successful than ours. Moreover, he would take it as his task to explain where their visual modules go wrong, to explain why their visual modules get it wrong when they get it wrong. Therefore, I conclude that Marr, in line with commonsense, would attribute the content *square* to state S1 in both Edgar and his twin.

#### 4.5 An externalist rejoinder

There is an important line of response according to which the externalist can concede the point that sameness of behaviour implies sameness of visual content whilst developing the circle-square case in such a way as to generate a genuine counterexample to individualism. The argument in question is due to Davies (1992) and runs thus. It is a mistake to think that twins that inhabit different environments must make just the same bodily movements. If twins live in environments that differ in terms of gravity, or in terms of the density of the medium through which they move, then identical nervous impulses and muscle contractions will issue in subtly different bodily movements. Suppose that the respective environments of Edgar and his twin differ in this kind of way. The very internal phenomena that cause Edgar to trace the shape of a square in the air cause Twin-Edgar to trace the shape of a circle and the very internal phenomena that cause Edgar to walk in a square-wise fashion cause Twin-Edgar to walk in a circle. Hence it is true of both twins that their behaviour constitutes a well adapted response to its distal cause. And thus the behavioural facts do not tell against

---

there being any breakdown in the relationship between visual state and behaviour. After all, there is no reason why a community couldn't adopt the convention of representing circular-shaped objects by means of squares. Whether or not Twin-Edgar's behaviour could plausibly be interpreted in this light, the response breaks down when it comes to circle-caused behaviour, the primary purpose of which is not to communicate the subject's conception of the shape of whatever object is impinging upon him. The way in which Twin-Edgar handles and circumnavigates circular-shaped objects is so inappropriate that there is no way of avoiding concluding that he sees them as squares, save accepting that there is a breakdown in the relationship between his visual states and his behaviour.

the attribution of the content *square* to S1 in Edgar and *circle* to the same state in his twin.

Ingenious though this argument is, I think it fails, as it overlooks an important fact about the way our bodies work. When I make a movement I am usually aware of the nature of the movement that I have made. For example, if I trace a square in the air with my hand I will be aware that I have so behaved. This awareness is a form of conscious experience (I have a sensation of tracing a square with my hand), but does not rely on vision; even with my eyes closed I am sensitive to the way in which I move my limbs. Having this capacity to monitor my movements plays an important role in enabling me to interact successfully with the world. Underlying this capacity are mechanisms within my body that provide my brain with information that enables it to determine the extent, direction and velocity of the movement of my limbs as I behave. These mechanisms will be sensitive not to (or not just to) the nature of the muscular contractions and nervous impulses that cause behaviour, but to such factors as the degree and direction of bone movement within sockets at the joints, and phenomena issuing from these movements such as the tension of ligaments and tendons. When I trace a square with my hand, the activity at my joints is subtly different from that that takes place when I trace a circle. This difference manifests itself in the feedback that my brain receives, which in turn results in my having quite different sensations as to the way in which I have moved. All this is true not just of me but of all humans and, of course, of Twin-Edgar and his fellows.

I concede that identical internal events (muscle contractions, nervous impulses and the like) could cause different bodily movements in twins that inhabit divergent environments. But surely the feedback that these movements generate will differ from one twin to the next. Edgar traces a square with his hand, whereas Twin-Edgar traces a circle. This difference in their respective bodily movements will be reflected in the activity that takes place at their joints; for example, there will be subtle differences in the direction and extent of the motion of bones within the sockets that link them to adjoining bones. If there were no such differences, the twin's, being twins, would surely make identical bodily movements. Given that the feedback mechanisms are sensitive precisely to such

phenomena, the response of Edgar's mechanism will be intrinsically different from that of his twins corresponding mechanism. Consequently, their subsequent experiences of how they moved will diverge; Edgar will have the experience of tracing a square, whereas Twin-Edgar will have the sort of experience (qualitatively and physically) that Edgar normally has when he traces a circle. The important point is that whatever contents we attribute to the internal feedback states of the twins, the fact is that they will differ at the physical level; that is, they will not be twins. Thus Davies has sketched a highly implausible scenario; twins with bodies that work anything like the way in which our bodies work cannot make different bodily movements. If two individuals move in different ways then they cannot be twins.

Another point worth making is that all this implies that Twin-Edgar will be woefully ill-adapted to his home environment. Suppose that I find myself in an environment like Twin-Edgar's. I see a circular-shaped object which causes a tokening of state S1 (a state which, according to the externalist, has the content *square*) and a visual experience *as of* a square. I attempt to behave towards this object in a square-wise fashion, but end up behaving as if it were a circle. With my eyes I will see myself engaging in "square" behaviour but from the inside it will feel like I am behaving in a "circular" fashion. In other words, the information that my brain receives from the visual module will clash with that that it receives from the internal feedback mechanism, and this clash will manifest itself in my being befuddled, experiencing a sense of disquiet, and coming to believe that something has drastically gone wrong. Clearly I would have problems leading a life in this environment if I had much contact with circular and square-shaped objects. To avoid grinding to a halt in a state of utter confusion, I would have to learn to ignore an important source of information as to the nature of my bodily movements. And it is difficult to see how a creature like me could have evolved in such an environment.

Twin-Edgar is going to face the same problems in his home environment that I do when I visit. When he is confronted with a circular-shaped object and acts "naturally", he is going to receive mixed messages concerning how he has behaved, mixed messages that will drive him into a state of confusion and a general feeling



that something is going wrong. This will be the case however we describe the contents of his internal states. The only way that Twin-Edgar will be able to overcome this problem will be by learning to ignore either what his eyes tell him in such situations, or what his internal feedback mechanism tells him. But if he has to undergo such a learning procedure he can hardly be well adapted to his environment; it is difficult to see how evolutionary processes could have cooked up a creature like him. And if Twin-Edgar isn't well adapted to his home environment, then he is of not much use to the externalist cause, for what the externalist ideally wants is a case of twins who inhabit radically different environments to which they are equally well adapted. The less well adapted a creature is, the less compelling are the grounds for regarding it as correctly representing its distal stimuli, and thus the less compelling are the grounds for thinking that the nature of its environment is reflected in the contents of its visual states.

I therefore conclude that Davies' attempt to modify the square-circle case fails. In this case, just as in the shadow-crack case, Marr would attribute just the same contents to the corresponding visual states of the twins in question. I think that this result is grounds for scepticism that there are any plausible counterexamples to individualism; it is difficult to see how there could be a plausible case of well adapted twins who inhabit environments that diverge to such an extent as to tell against the attribution of identical contents.

#### **4.6 Behaviour versus environment**

My argument in the preceeding section relies on the idea that Marr places a great reliance on the behaviour of the subjects whose visual modules he studies. Such a portrayal of Marr's approach might appear to be in tension with the way in which I described his theory at the beginning of this section (and, indeed, with my general account of scientific psychological explanation); for I stressed both that Marr proceeded by reflecting on the nature of the extra-cranial world, and that appeals to facts about that world play a significant role in his theory. Fortunately, this apparent tension can be resolved, and can be resolved in such a way as to undermine Burge's initial externalist argument.



As we have seen, scientific psychologists regard our perceptual and cognitive modules as information-processing systems, systems that generate information from information. To describe a system as generating information from information is to say more than that it generates meaningful symbols from meaningful symbols. In addition, it is to imply that the system generates symbols that correctly represent facts about a certain subject matter from symbols that correctly represent facts about some other subject matter. In other words, it is to imply a good deal of success. Thus scientific psychology operates with the idea that our perceptual and cognitive modules are largely successful. In saying this, I do not contradict any of my earlier assertions, as this assumption of success is based upon behavioural evidence and in no way commits the psychologist to the view that cognitive and perceptual modules must by definition be successful; it's just that, in point of empirical fact, our modules are largely reliable.

When a system generates information from information, that it does so, or that it is able to do so, by the means that it employs, has a lot to do with the nature of the world. In general, what information can be generated from what information (and how) is determined by the way the world is. For example, had the world been suitably different I would not be able to extract information about my current bank balance from information about its level at a previous date and all my intervening banking transactions. Thus it is important to highlight the relevant facts about the world in explaining my success; in explaining how I manage to reach true conclusions about how much money I have in my bank account. The same applies to perceptual and cognitive modules; if they are successful, if they do generate information from information, then an explanation of that success must appeal to the relevant (contingent) facts about the world. But none of this implies that the nature of the world external to an information-processing system determines the content of the representations that it manipulates. Suppose that the content of our visual states was individualistic, or was assumed to be so by scientific psychology. It would still be necessary to appeal to relevant facts about the world in order to explain the success of the visual module in generating information or veridical representations, as that success would partly be the product of such external facts. Hence the

fact that scientific psychology (and Marr) appeals to facts about the extra cranial-world does not imply that it (or he) individuates content non-individualistically.

It is true that Marr reflects on features of the world in order to determine how the visual module works, and that there is an important respect in which the contents that he attributes to our visual states is environment driven. But once again this does not tell against his being an individualist. Suppose that you were confronted by a system whose observable behaviour indicated that it solved a certain information-processing problem. What would be the most effective way of determining how the system solved that problem, of explaining its success? A very sensible strategy would be to reflect on the system's environment in order to work out what potential solutions that environment permits, and which it rules out. If a certain solution is discovered to be theoretically possible in that environment, then a hypothesis worthy of consideration is that the system solves the problem in that way. (After all, in any environment there will be only a small number of ways to solve any complex information-processing problem). This hypothesis can then be checked against behavioural evidence. If the behavioural evidence supports the hypothesis then it can be tentatively endorsed, an endorsement which involves attributing certain contents to the system's internal states. If this strategy is pursued, then an explanation will be constructed by means of a consideration of the world external to the system. And the contents attributed to the system's internal states will be environment driven in the sense that had the environment been significantly different, then a different explanation would have been generated, and thus different contents attributed.

To adopt this strategy is not thereby to individuate the contents one attributes to the system in question non-individualistically. Someone who adopted the strategy and endorsed the resultant explanation would be free to regard the internal states of all the system's twins as having just the same content no matter what their home environment was like. It's just that she would be forced to concede that some of those twins would systematically generate false answers to the information-processing problem that they attempted to solve.

The strategy that I have described is a simplified version of that employed by Marr. It involves considering both behaviour and the extra-cranial environment. Consideration of behaviour helps to determine the extent of the visual module's success, and to choose between alternative candidate explanations of that success. And consideration of the environment helps to generate such candidate explanations, and reveal why the visual module's operations work to the extent that they do.<sup>75</sup> Thus, by parity of reasoning, Marr's approach is consistent with his individuating visual content individualistically; that he examines the extra-cranial world, and that his content attributions are environment driven, merely reflects the fact that he (on the basis of an observation of our behaviour) is of the opinion that the human visual module is largely successful in generating veridical representations of the external world from retinal images.

This, I take it, resolves the tension between, on the one hand, the claim that Marr is sensitive to behaviour in the way that I have described and, on the other hand, the claim that a consideration of the environment and an appeal to environmental factors plays a fundamental role in Marr's approach, and theory. It also serves to undermine Burge's argument by indicating how the success-orientation of Marr's approach and the fact that his content attributions are environment driven, is consistent with his individuating visual content individualistically.

---

<sup>75</sup> In a little more detail: Human behaviour suggests to Marr that we generally correctly represent such properties of objects as their shape, size, colour, surface markings, motion, and the like: hence that behaviour indicates that the visual module succeeds in working out the shape, size, colour, surface markings, motion, and the like, of distal stimuli from the retinal images that it takes as input. Given that the visual module's ability to do this partly depends on the nature of the extra-cranial world, Marr reflects on the nature of that world for cues as to just how it does it. Candidate explanations generated in this way can then be checked off against behavioural evidence. For example, our performance in the task of categorising objects on the basis of stick figure representations of them supports the idea that the visual module represents the shape of objects as, or by means of, generalised cones or collections of generalised cones.

#### 4.7 Conclusion

I therefore conclude that the case against Marr's being an individualist is unconvincing. Neither the crack-shadow nor the circle-square case can be developed in such a way as to tell against individualism. Moreover, it is consistent with the success-orientation of Marr's theory, and the fact that he appeals to contingent facts about our environment, that Marrian contents are locally supervenient. However, that is not to say that Marr employs a notion of narrow content in the sense of a type of content divergent from that employed by ordinary folk (or one that is unfamiliar from the folk perspective). Of course some of the representations that figure in Marr's theory (the primal sketch and the  $2\frac{1}{2}$ -D sketch, for example) are unfamiliar. But many of the properties that these representations express, and thus the contents that they have, are familiar enough. After all, what is so strange about a representation representing an object as being square-shaped or as having a small dark mark on its surface? This point hints at a way of reconciling the conclusion of this chapter with the line I developed in Chapter 3.

In Chapter 3 I argued that due to the practical problems of "going narrow" scientific psychology should not employ a notion of narrow content if at all possible, and suggested that it is indeed possible to avoid narrow content. I also attempted to undermine various pro-individualist arguments, and claimed that the contents attributed by psychologists in the course of accounting for some of our recognitional and classificatory capacities are not locally supervenient. It is natural to think that all of this is in tension with the conclusion of the present chapter. However, appearances notwithstanding, there is a harmony to be found.

A first point by way of reconciliation is this. To say that some of the contents that scientific psychology attributes to our psychological states are locally supervenient is not to imply that they all are. Our cognitive capacities are many and varied, so it should perhaps come as no surprise that the contents of the symbols manipulated by some of our cognitive modules are locally supervenient, whereas those of the symbols manipulated by others are not.

The second, and most important point, runs thus. In Chapter 3 I emphasised the continuity between folk psychology and its scientific cousin. In the case of Marr's theory I think that such a continuity also holds. The contents that Marr attributes to the states of the visual module are very much like the contents that folk psychologists routinely attribute to our personal-level visual states. It is not just that there is an overlap in the properties that the respective states express. In addition, as far as folk psychology is concerned, the contents of our visual states are locally supervenient; environmental differences and differences in typical distal cause do not make for differences in content in the way in which they do with thoughts involving natural kind concepts. Therefore, in both this chapter and Chapter 3, I have been effectively arguing that there is no fundamental break between the intentional properties attributed to our psychological states by folk psychology on the one hand, and scientific psychology, on the other.



## Chapter 5

# Causal Powers and a Metaphysical Argument for Individualism

### 5.1 Introduction

I have been arguing for a tentative externalism: scientific psychology need not employ a notion of narrow content at odds with folk psychological content. It can legitimately, and perhaps sometimes should, appeal to intentional properties that are not locally supervenient. In arguing for this conclusion I have appealed to practical considerations and to actual psychological practice. This raises the question of whether my position falls foul of Fodor's metaphysical arguments for individualism. In this chapter I will address this question and attempt to establish a negative answer; Fodor gives us no compelling reasons for thinking that scientific psychology must respect the local supervenience of the psychological upon the physical.

### 5.2 The argument from causal powers

In Chapter 2 of *Psychosemantics* Fodor presents what he describes as a metaphysical argument for the conclusion that scientific psychology must individuate the states and events that figure in its theories and explanations in such a way as to respect the supervenience of the psychological on the neurophysiological. Given the reasonable assumption that physical twins are neurophysiologically identical, this implies that physical twins are psychological duplicates; in other words, that psychological properties are locally supervenient. For reasons that will soon become apparent, I will call this argument the argument from causal powers. It rests upon a claim about the nature of science and can be described in the following terms.

The aim of science is to construct causal explanations. Constructing a causal explanation involves subsuming events under a causal

generalisation. Such generalisations 'subsume the things they apply to in virtue of the causal properties of the things they apply to' (*Psychosemantics* p.34). This fact has consequences for the way that sciences individuate the states, events, and entities that feature in their explanations, for they must individuate them in terms of their causal powers. Thus a science will allot distinct states, events, and entities to the same type only if they are relevantly similar in their causal powers, and will distinguish between those that diverge in their causal powers. And a property can figure in a scientific taxonomy only if it affects the causal powers of whatever has it.<sup>76</sup> Consequently, scientific psychology must individuate in terms of causal powers. This entails that it must individuate in a manner that respects the local supervenience of the psychological on the physical as the psychological states of physical twins are equivalent in their causal powers; such properties as distinguish between twins do not affect causal powers and therefore have no place in a scientific taxonomy.

As it stands, this argument is hardly conclusive. On the one hand, one might wish to question the claim that science must always individuate in terms of causal powers. And on the other, one might wish to argue that the causal powers of psychological states are not locally supervenient, so that twins could have psychological states

---

<sup>76</sup> Fodor expresses this putative fact about scientific taxonomy by saying that sciences are individualistic. His use of the term "individualism" is potentially misleading as it is at odds with the standard usage. In Fodor's terminology, to individuate individualistically is not thereby to individuate non-relationally or solipsistically, as relational properties sometimes affect the causal powers of whatever has them, and thus are fit to figure in a scientific taxonomy. An example of a relational property that features in a scientific taxonomy in virtue of suitably affecting the causal powers of whatever has it is the property of being a planet. As Fodor puts it:

'being a planet' is a relational property par excellence, but it's one that individualism permits to operate in astronomical taxonomy. For whether you are a planet affects your trajectory, and your trajectory determines what you bump into; so whether you're a planet affects your causal powers, which is all that individualism asks for. Equivalently, there are causal laws that things satisfy in virtue of being planets. (*Psychosemantics*, p. 43)

that diverged in their causal powers (and thus belong to different psychological types). Fodor invests a good deal of effort into defeating the second of these potential lines of objection, and it is here that the power and interest of his argument lies. But before examining that line of his thought I will make a couple of remarks in the spirit of the first potential line of attack.

Fodor accepts that folk psychology does not individuate in terms of causal powers; there is a mismatch between the folk and scientific psychological taxonomy. But he stresses that the primary business of folk psychology is to produce causal explanations, and that in constructing such explanations its practitioners utilise causal generalisations relating folk psychological states and behaviour. Moreover, he thinks that the states that folk psychologists appeal to have reality, and that the generalisations they employ and the explanations they construct are often true. What this suggests is that the demands of causal explanation can be met without the utilisation of a taxonomy that individuates in terms of causal powers. Thus, one might ask, why need science in general, and scientific psychology in particular, employ an individualistic taxonomic scheme? Folk psychology gets by without doing so. In short, there is a tension between Fodor's pronouncements on the nature of scientific taxonomy and his account of, and enthusiasm for, folk psychology. A claim that Fodor makes in an endnote perhaps provides something of an answer to this charge. He says that the folk psychological ascription of intentional states often serves purposes other than causal explanation, and this is reflected in its taxonomy; hence the difference between it and the taxonomy of a discipline whose only concern is to construct causal explanations.

One reason why you might want to know what Psmith believes is in order to predict how he will behave. But another reason is that beliefs are often true, so that if you know what Psmith believes, you have some basis for inferring how the world is. The relevant property of Psmith's belief for this latter purpose, however, is not their causal powers but something like *what information they transmit . . .* And, quite generally, what information a thing transmits depends on relational properties of the thing which may not affect its causal powers. (1987, p. 157).

A couple of points in response to this. First, Fodor effectively concedes that a taxonomy, namely that of folk psychology, can work for the purposes of causal explanation even though, strictly speaking, it is not a taxonomy in terms of causal powers. So why can't scientific psychology employ the same taxonomy? Surely it would work for the purposes of scientific psychological explanation. Even if scientific psychologists are not interested in reading off facts about how the world is from the states of their subjects, there would be an advantage in holding onto the folk taxonomy. The advantage is that scientific psychology would be somewhat easier to engage in if it utilised the familiar notion of content than it would be if it employed a notion of narrow content. This is not just because formulating a workable notion of narrow content is so difficult, hence making it very hard for psychologists to work out just what narrow states their subjects are in. It is also because we are so steeped in folk psychology that we would have to wrestle with our instincts to see and categorise the psychological states of our fellows in a way that conflicted with the way we, qua folk psychologists, see and categorise the psychological states of our fellows.<sup>77</sup> Put crudely the point is this: why go to all the bother (and it would be a lot of bother) of constructing and employing a psychological taxonomy at variance with that of folk psychology if the latter will do for the purposes of the causal explanation of such psychological events as behaviour and the tokening of psychological states?<sup>78</sup>

---

<sup>77</sup> Psychologists would have to wrestle with their instincts quite literally if Fodor is right in thinking that much of folk psychology is innate. (He advances, and argues for, this innateness hypothesis in 'The Present State of the Innateness Controversy').

<sup>78</sup> In Chapter 2 I argued that there are practical reasons for avoiding "going narrow", reasons that may well outweigh the benefits of abandoning folk psychological content if twins are thin on the ground. The present point is very much in the same spirit. If the folk psychological taxonomy will work for all practical purposes then it would be foolish to abandon it just for the sake of satisfying some ideal of pure science. Whether or not the practical drawbacks of abandoning the folk psychological taxonomy outweigh the benefits is, I take it, an open empirical question that has a lot to do with whether we have any twins who (from the folk psychological perspective) psychologically diverge from us.

Second, Fodor's point that the information that a state carries depends on relational properties that do not affect causal powers suggests the idea that if the account of the nature of scientific psychology that I developed in Chapter 2 is correct, scientific psychology may well have to employ a taxonomy that recognised properties that do not affect causal powers. This is because, on my account, scientific psychology must appeal to the information that psychological states carry and that psychological processes extract or work out in order to account for our cognitive capacities. In short, the plausibility of Fodor's line on the nature of the taxonomy of scientific psychology might be closely bound to the plausibility of his view that scientific psychology is primarily in the business of constructing singular causal explanations of behaviour and of the tokening of psychological states, explanations that fit the deductive nomological model.

### **5.3 A putative counterexample**

For now let's assume that Fodor is right in claiming that scientific psychology must individuate in terms of causal powers. The problem is that it isn't obviously the case that such a psychology will respect local supervenience of the psychological on the physical. So the question is: can Fodor rule out the possibility of physical twins whose psychological states do not agree in their causal powers? It is to this question that I now turn.

The Putnamian twins might appear to constitute a problem for Fodor for the following reason. Suppose Oscar has a thought which causes him to say "fetch me a glass of water". Given that he lives here on Earth, this utterance, and thus the thought that caused it, results in him being presented with a glass of water. The corresponding thought of Oscar<sub>2</sub> likewise causes him to utter the sentence "fetch me a glass of water", but due to his living on Twin Earth the effects of this utterance, and the thought that caused it, is the presentation, not of a glass of water, but of a glass of twater. Therefore, so the argument concludes, the respective thoughts differ in their causal powers and thus belong to different psychological types.



In responding to such putative counterexamples to his supervenience thesis, Fodor argues that 'identity of causal powers has to be assessed *across* contexts, not *within* contexts' (p.35). Thus determining whether two distinct states have the same causal powers involves considering not merely what effects they actually bring about but what effects they would bring about were they located in different contexts. When we take into account the relevant counterfactuals we see that the thoughts of the twins have the same causal powers; for were Oscar<sub>2</sub> located here on Earth (that is in Oscar's context) his thought would land him with water and were Oscar located on Twin Earth (that is in Oscar<sub>2</sub>'s context) his thought would land him with twater.

This reply is decisive, but the Putnamian twins present Fodor with a far more substantial problem. Here is one form of the problem. Psychological states have behavioural and psychological consequences; that is, they cause both behaviour and the tokening of other psychological states. Such effects have intentional properties, properties that they inherit from their causes. The effects of Oscar's water thoughts differ in their intentional properties from the corresponding effects of Oscar<sub>2</sub>'s twater thoughts. For example, the behaviour that Oscar's desire for water causes is water-seeking behaviour, whereas the physically identical behaviour that Oscar<sub>2</sub>'s desire for twater causes is twater-seeking behaviour. What this suggests is that Oscar's water thoughts diverge in their causal powers from Oscar<sub>2</sub>'s twater thoughts, as the former have the power to cause water behaviour and water thoughts, a power that the latter lack. Similarly Oscar<sub>2</sub>'s twater thoughts have a power Oscar's water thoughts lack, namely the power to cause twater behaviour and twater thoughts.

Application of the cross context test does nothing to unseat this conclusion, for if Oscar is transported to Twin Earth his water thoughts remain water thoughts and the behaviour that they cause maintains the property of being water behaviour. The corresponding point can be made about a transported Oscar<sub>2</sub>. As scientific psychology is concerned with explaining behaviour and the tokening of thoughts, it seems reasonable to conclude that this difference of causal powers is one that scientific psychology should care about, and thus that scientific psychology should assign Oscar's water thoughts

to a different psychological type than Oscar<sup>2</sup>'s corresponding twater thoughts.

Fodor will have none of this, and puts a lot of effort into defeating this putative counterexample to his supervenience claim. Essentially he has two major arguments for the conclusion that the psychological states of the Putnamian twins agree in their causal powers. The first appears in *Psychosemantics*, and is to the effect that to regard the thoughts of the twins as having different causal powers is to postulate crazy causal mechanisms and impossible causal laws. The second appears in "A modal argument for narrow content", and is to the effect that the difference between the twins and their respective thoughts fails to meet a necessary condition for their having different causal powers. This is because the difference between the twins' respective thoughts is conceptually related to the difference between their respective behaviours. Both these arguments will be examined (and rejected) in due course, but first of all I will examine a couple of preliminary (and less significant) arguments that Fodor presents, arguments that are supposed to discredit the idea that the twins' mental states differ in their causal powers.

The first of these two arguments draws a parallel between a pair of properties that clearly don't count in any scientific taxonomy and the properties that Fodor's externalist opponent thinks should figure in a scientific psychological taxonomy. The properties in question are that of being an H-particle and that of being a T-particle. A physical particle is an H-particle at some point in time if and only if Fodor's dime is heads up at that point in time. Similarly, a physical particle is a T-particle at a given point in time if and only if Fodor's dime is tails up at that time. Clearly, thinks Fodor, the property of being a T-particle and that of being an H-particle should not figure in the taxonomy of physics, as these properties do not affect the causal powers of whatever has them. Consider a particle P at time t<sub>1</sub>. Fodor's dime is heads up at time t<sub>1</sub>. At time t<sub>2</sub> Fodor flips his dime so that it becomes tails up. Clearly P's causal powers haven't changed from t<sub>1</sub> to t<sub>2</sub> solely in virtue of the change in the orientation of Fodor's dime. Yet despite the absurdity of the claim that these properties affect causal powers, one could construct an argument to the effect that they do in fact affect causal powers, an argument that

has just the same form as the above argument for the conclusion that the thoughts of the Putnamian twins differ in their causal powers.<sup>79</sup> If the latter argument works then so does the former, in virtue of their similarity in form. But the former argument can't work as it has an absurd conclusion. Therefore the latter argument must also be spurious.

In reply to this argument, it is unfair to draw a parallel between the properties of being an H (or a T) -particle and those properties that differentiate the twins. The extent of the differences between these pairs of properties will become evident in my discussion of Fodor's second main argument below. In the meantime suffice it to say the following. For a particle to be an H-particle there need be no causal relationship between it and Fodor's dime. But for a thought to be a water thought there must be some special and significant causal relationship between that thought, or the subject who has it, and water, for it is in virtue of not having the right causal relationships to water - or not having a history in which certain causal interactions with water figure - that Oscar<sub>2</sub>'s thoughts are not water thoughts. This kind of difference between the property of being a water thought (/water thought) and that of being an H-particle (/T-particle) suggests that one could admit the former to a scientific taxonomy without thereby admitting the latter.

A second reply involves accepting that the property of being an H-particle (and that of being a T-particle) affects causal powers. This doesn't automatically have the consequence that these properties should figure in the taxonomy of physics (or any other science), as to so figure a property must not just affect causal powers but must affect them in relevant ways. Perhaps physics should not distinguish between particle P at t<sub>1</sub> (when it is an H-particle) and P at time t<sub>2</sub>, not because it is identical in its causal powers at these two points in time, but because it is not *relevantly* different in its causal powers.

This brings us to Fodor's second attempt to discredit the idea that the thoughts of the twins diverge in their causal powers. According to this argument, to endorse this idea lands one with the absurd consequence that the twins are neurophysiologically distinct in

---

<sup>79</sup> The argument in question runs thus: 'Being H rather than being T does affect causal powers after all; for H-particles enter into H-particle interactions, and no T-particle does' (1987, p. 38).

virtue of there being differences between the causal powers of their respective brain states. For example, Oscar's brain states have the power to cause water behaviour, a power that the corresponding states of his twin do not.

This is not a very convincing argument. The obvious reply involves repeating the point that for science it is not causal powers per se that count for individuation, but relevant causal powers. Not caring about intentional properties, neurophysiology will not distinguish between brain states whose difference in causal powers is restricted to the intentional realm. Rather the brain states must diverge in their powers to produce effects individuated in terms of their neurophysiological properties. In a different context Fodor as good as makes the same point when he says:

sciences are forever cross-cutting one another's taxonomies. Chemistry doesn't care about the distinction between rivers and lakes; but geology does. Physics doesn't care about the distinction between bankers and butchers; but sociology does. (For that matter, physics doesn't care about the distinction between the Sun and Alpha Centauri either; sublime indifference!) None of this is surprising; things in Nature overlap in their causal powers to various degrees and in various respects; the sciences plat these overlaps, each in its own way. (1987, p. 45).

So much for Fodor's minor preliminary arguments against the claim that the psychological states of the Putnamian twins diverge in their causal powers. In the next section I will address the first of his main arguments.

#### 5.4 Crazy causal mechanisms and impossible laws

In *Psychosemantics* the primary argument against the claim that the Putnamian twins constitute a counterexample to the thesis that the causal powers of mental states are locally supervenient runs thus. If the properties of being an H-particle and being a T-particle affected the causal powers of whatever particle had them, then the causal powers of a particle would depend on the orientation of Fodor's dime. For the causal powers of a particle to depend upon the



orientation of Fodor's dime, there would have to be a causal mechanism or a fundamental law of nature to mediate the dependency. But there aren't any such mechanisms or laws, for if there were, Fodor's dime would be able to causally influence every particle in the universe, something that it clearly cannot do. In short, to claim that the property of being an H-particle and that of being a T-particle affects causal powers is to postulate "crazy causal mechanisms" or "impossible laws". To hold that the psychological states of the twins differed in their causal powers would similarly be to postulate crazy causal mechanisms or impossible causal laws. If the psychological states of the twins differed in their causal powers, then the causal powers of such states would depend upon the character of the environment in which the individual who had them existed. For there to be such a dependency relationship there would have to be some mediating mechanism, a mechanism that enabled the character of an individual's environment to affect the causal powers of his psychological states without affecting his physiology.

But there is no such mechanism; you *can't* affect the causal powers of a person's mental states without affecting his physiology. That's not a conceptual claim or a metaphysical claim, of course. It's a contingent fact about how God made the world. God made the world such that the mechanisms by which environmental variables affect organic behaviours run via their effects on the organism's nervous system, or so, at least, all the physiologists I know assure me. (1987, p. 40).

How should we respond to this argument? Fodor's thought (as evidenced by the above quoted passage) seems to be that it is a mistake to hold that there is a dependency relationship between the character of an individual's environment and the causal powers of his psychological states, because such a view leads to the absurd conclusion that changes in an individual's environment can cause changes in the causal powers of his psychological states without having any effects on his physiology. A first problem with this is that the supposedly absurd upshot is not obviously absurd. Certainly there are plenty of examples from outside psychology where the causal powers of a thing change due to changes in the environment



without the entity in question undergoing any internal physical changes. For example, I have the power to lift Fang. That I have this power depends partly on the character of the world external to me, in particular on how heavy Fang is. Had Fang been twice his weight then, given how I am internally, I wouldn't have the power to lift him. The world could change in such a way that, without causing any internal changes in me, I lost my power to lift Fang and thus underwent a change in my causal powers. This would happen if Fang underwent a spurt of growth that pushed him to twice his present weight.<sup>80</sup> Perhaps there is a psychological analogue to such a case.

Causal powers are powers to produce effects in the world. These effects are often at some distance from the power-bearing individual in question. Moreover, they are often described and individuated in terms of relational properties that are fixed by distant facts in the world. Consequently, it is not at all surprising that a thing's causal powers could change without it undergoing any intrinsic physical change; for the change in causal powers could be brought about by changes in the external world. To say this is not to make such changes in causal powers miraculous, mysterious or inexplicable. Nor is it to sever the connection between the physical and the causal powers of things, for the kinds of distant changes, and changes in relational properties, that can affect causal powers will always be the products of physical changes in the world. For example, my losing the power to lift Fang is a result of a physical change in the world, namely, a change in his size and weight, even though it is not the result of a change in my intrinsic physical properties.

---

<sup>80</sup> It might be objected that my causal powers haven't changed at all in this case as the maximum weight that I can lift will have remained unchanged. In response, I would say that certain of my causal powers have remained unchanged, but my powers with respect to lifting Fang are not among them. How we describe and individuate effects will influence how we describe and individuate causal powers. And how we describe and individuate effects is an interest relative matter. Given my interests and purposes it makes good practical sense to describe and individuate effects in terms of the individual objects involved so that my lifting Fang at  $t_1$  is an event of the same type as my lifting him at  $t_2$ , despite the fact that his size and weight has changed between these two points in time. Thus, when Fang grows too heavy for me, I can no longer cause an effect of a type that I could once bring about with ease.

It is important to realise that to hold that there could be a psychological analogue of the example of my losing the power to lift Fang isn't to contradict what Fodor's physiologist friends assure him. Of course an environmental event can't cause a behavioural event without affecting the individual in question's physiology. But the events the possibility of which the externalist envisages are events of a different kind from such behavioural events, for they are constituted by a stable, pre-existing state's undergoing a change in certain of its dispositional properties. If this sounds unconvincing consider my power to lift Fang. Intuitively the event of my lifting Fang is a fundamentally different kind of event from that of my losing my power to lift Fang. The latter event, unlike the former, involves a change in my dispositional properties and can take place without my undergoing any internal physical change. The physiologist does not mean to rule out the possibility of such events, but only the possibility of events of the first sort. In other words, the physiologist only means to rule out the possibility of such events as that of my lifting Fang in response to an event in the environment without my undergoing any internal physiological change.

A second objection to Fodor's argument is that to view the psychological states of the twins as diverging in their causal powers is not thereby to commit oneself to the allegedly absurd consequence that he describes. Davies expresses the point in the following manner.

Now, a typical consequence of externalism is that, if a neurophysiological twin of an actual subject *had* been set in a different environment then our actual taxonomy would not have applied in the counterfactually imagined environment to classify the twin in the same way as the actual subject is classified. In that sense, he *would have been* psychologically different from the actual subject. But it does not follow from this that a way of *making* the actual subject psychologically different in those ways is by *changing* his environment now. Still less does it follow from the externalist claim about the counterfactual environment, that we can *make* the subject psychologically different now without making any physiological difference. (1986, pp. 272-3)

To see this consider the case of the Putnamian twins. Oscar's water thoughts are water thoughts (as opposed to twater thoughts) not just because he lives in a watery world, but because he stands in a certain complex causal relationship to water, or because he has a history in which certain causal interactions with water figure. The standard intuition is that these causal relationships are such that they would not be overridden were Oscar transported to Twin Earth.<sup>81</sup> His being on Twin Earth and interacting with twater wouldn't be enough to make any of his thoughts twater thoughts. Similarly, were Oscar<sup>2</sup> transported to Earth, the thoughts resultant of his interactions with water, and those underlying his utterance of sentences containing the word "water", would not be water thoughts; rather they would be twater thoughts.

However, a qualification is needed here. The standard intuition also has it that were a transported Oscar to hang around Twin Earth long enough, his "water" thoughts would eventually become twater thoughts for he would eventually become embedded in a watery world and a linguistic community that used the word "water" to mean *twater*. This raises the possibility of a transplanted Oscar being neurophysiologically identical to one of his previous selves yet having thoughts that diverged in their causal powers from their earlier counterparts. But this isn't quite an instance of the "absurd" consequence that Fodor accuses the externalist as being committed to. For in the period intervening between Oscar's having water thoughts and his having twater thoughts there is an awful lot of causal interaction between Oscar and the Twin Earth environment (and in particular between him and the linguistic community on Twin Earth). Without such causal interaction he would not have become sufficiently embedded in the Twin Earth environment to be capable of having any twater thoughts. Thus this case is very different to that where a particle changes from being an H-particle to being a T-

---

<sup>81</sup> This would appear to be accepted by Fodor, as is indicated when he says:

although I've heard it suggested that mental states construed nonindividually are easily bruised and don't 'travel', the contrary assumption would in fact seem to be secure. The standard intuition about 'visiting' cases is that if, standing on Twin Earth, I say "That's water" about a puddle of XYZ, then what I say is *false*. Which it wouldn't be if I were speaking English<sup>2</sup>.

particle (and thus from engaging in H-particle interactions to engaging in T-particle interactions) as a result of a change in the orientation of Fodor's dime. This is because in this case the change isn't the consequence of a complex history of causal interactions between the particle in question and Fodor's dime.

These two objections to Fodor's first major argument against the claim that the psychological states of the Putnamian twins diverge in their causal powers are hardly conclusive. However, I do think they serve to dent his argument somewhat. What is needed is a detailed examination of the nature of causal powers that does justice to the intuition that a thing's causal powers are determined by its intrinsic physical nature, yet explains that and how there can be differences between the causal powers of distinct things (and changes in a particular thing's causal powers) in the absence of intrinsic physical differences (and changes). This is a task that I will put off until section 5.6. In the meantime I will turn my attention to Fodor's second argument, an argument that appears in 'A Modal Argument for Narrow Content'.

### **5.5 A modal argument**

It will be helpful to recall the putative counterexample presented by the case of the Putnamian twins to Fodor's supervenience claim. It runs thus. Oscar has thoughts which have the property of being water thoughts in virtue of his causal relations to water (or his having a history in which certain significant causal interactions with water figure). The corresponding thoughts of Oscar<sub>2</sub> are not water thoughts, for Oscar<sub>2</sub> does not bear the appropriate causal relations to water; rather they are twater thoughts. In virtue of its etiology, the behaviour caused by Oscar's water thoughts is water behaviour, whereas that caused by Oscar<sub>2</sub>'s twater thoughts is not water behaviour but, rather, twater behaviour. A consequence of this difference between their respective behavioural effects is that Oscar's water thoughts have the power to cause water behaviour, a power that his twins twater thoughts do not have; and Oscar<sub>2</sub>'s twater thoughts have the power to cause twater behaviour, a power that his twins corresponding thoughts do not have. Given its concern with explaining behaviour under its intentional description, this

difference in causal powers is one that should be recognised by scientific psychology, hence Fodor's supervenience claim is mistaken.

It will also be helpful to distinguish a second, related version of the putative counterexample/objection. It runs thus. In virtue of the causal connections he bears to water, the thoughts that Oscar expresses with sentences containing the word "water" are water thoughts. Thus Oscar has the power to think water thoughts. This is a causal power that Oscar<sub>2</sub> does not have, as the thoughts that he expresses with sentences containing the word "water" are twater thoughts. What he has is the power to think twater thoughts. Thus the twins differ in their causal powers, and given that psychology is concerned with explaining behaviour by reference to its intentional causes, this is a difference in causal powers that should be recognised by scientific psychology.

In 'A Modal Argument for Narrow Content', Fodor responds to these putative counterexamples.<sup>82</sup> He begins by describing a schema which both cases fit, and naming it "schema S". Schema S is as follows. C1 and C2 are a pair of causes, and E1 and E2 are their respective effects:

C1 differs from C2 in that C1 has cause property CP1 where C2 has cause property CP2.

E1 differs from E2 in that E1 has effect property EP1 and E2 has effect property EP2.

The difference between C1 and C2 is responsible for the difference between E1 and E2 in the sense that, if C1 had had CP2 rather than CP1, then E1 would have had EP2 rather than EP1; and if C2 had had CP1 rather than CP2, E2 would have had EP1 rather than EP2. (1991, p. 9).

The first of the putative counterexamples fits schema S in this way: C1 is a thought of Oscar and C2 is a corresponding thought of his twin. CP1 is the property of being a water thought and CP2 the

---

<sup>82</sup> 'A Modal Argument for Narrow Content' is a highly technical paper. Consequently, my discussion of it is also highly technical, especially by my standards. For some different, but equally technical, responses to Fodor's reasoning see Peacocke (1994), and Baker (1995).



property of being a twater thought. E1 and E2 are instances of behaviour of Oscar and his twin respectively. EP1 is the property of being water behaviour and EP2 is the property of being twater behaviour.

And the second putative counterexample fits S in this way: C1 is Oscar and C2 Oscar2. CP1 is the property of being causally connected to water and CP2 is the property of being causally connected to twater. E1 and E2 are Oscar's and his twin's corresponding thoughts respectively. EP1 is the property of being a water thought and EP2 is the property of being a twater thought.

Fodor argues that not every instance of schema S is a bona fide case of a divergence in causal powers. That is, not every instance of S is a case 'where the difference between having CP1 and having CP2 is a difference in causal power in virtue of its responsibility for the difference between E1 and E2'. This raises the question of whether the putative counterexamples are bona fide cases. He formulates a condition that he thinks that any instance of schema S of a certain type (a type that the putative counterexamples belong to; see footnote 6) must satisfy if it is to be bona fide. He then argues that the putative counterexamples fail to satisfy this condition, and thus the threat that they pose to the supervenience claim evaporates.<sup>83</sup>

The necessary condition that Fodor presents (a condition that he labels condition C) is essentially this: the difference between having CP1 and having CP2 is a difference in causal power in virtue of its

---

<sup>83</sup> It is important to get clear on the following point. In the case where an instance of S fails to meet the necessary condition, Fodor is not thereby committed to concluding that the causes in question agree in their causal powers. Rather, all he is committed to is rejecting the claim that C1 has different causal powers than C2 in virtue of CP1's responsibility for E1's having EP1 rather than EP2 (or in virtue of CP2's responsibility for E2's having EP2 rather than EP1). It may well be that CP1 (or CP2) is responsible for some other property of E1 (or E2) in virtue of which C1's causal powers diverge from those of C2. Or it might be that C1 (or C2) has some other property not shared by its counterpart which affects its causal powers. Indeed, argues Fodor, any two causes that differ in some contingent property will thereby diverge in their causal powers due to the possibility of constructing a machine for detecting that property. The cause that has the contingent property in question will have the power to cause a detector of that property to go into the positive state, a power which its counterpart (which doesn't have the property) will not have.

responsibility for the difference between E1's having EP1 and its having EP2 only when this difference between the effects is nonconceptually related to the difference between the causes. Thus if there is a conceptual relationship between having CP1 (rather than CP2) and having effects that have EP1 (rather than EP2) so that to have EP1 (rather than EP2) *just is* to be caused by something with CP1 (rather than CP2) then the necessary condition is not satisfied.<sup>84</sup>

Essentially, Fodor's justification for the condition is twofold. First, it is consistent with his intuitions in that the instances of S which he thinks are bona fide pass the test and those that he thinks are not fail the test. Thus we find him saying: 'My evidence for the acceptability of this condition will be largely that it sorts examples that I have just run through in an intuitively satisfactory way' (p. 12). And second, it coheres with Humean considerations about the nature of causation. Thus we find him saying: 'This . . . condition is motivated both by our intuitions about the examples and by the Humean consideration that causal powers are, after all, powers to enter into nonconceptual relations' (p.24). Given this, it would be a severe blow for Fodor if it could be shown that there is something wrong with his intuitions, and that there is no inconsistency between holding a Humean view of causation and regarding water and twater thoughts as diverging in their causal powers. I will now attempt to show that there is in fact something wrong with his intuitions. (An attempt to effect a reconciliation between externalism and a Humean view of causation will come later).

Fodor gives three examples of what he takes to be non bona fide instances of schema S. Firstly, there is the case where CP1 is the property of being C1, CP2 is the property of being C2, EP1 is the property of being the effect of C1 and EP2 is the property of being the effect of C2. Of this case he says 'it seems a priori obvious that this is

---

<sup>84</sup> An important point to note is that condition C is not supposed to be necessary for all instances of S. That is, there are instances of S that are bona fide despite the fact that they fail to satisfy C. In such cases 'the property that distinguishes the causes is itself the property of having a certain causal power' (p. 15). Such properties are noncontingently causal powers; that is, it is noncontingent that things that have such properties have certain causal powers in virtue of having them. An example of such a property is that of being soluble : to be soluble *just is* to have the power to dissolve when placed in water.

not a case where having CP1 is a causal power of C's in virtue of its responsibility for E's having EP1' (p. 11).

Fodor's mistake here is in thinking that the example is an instance of schema S. As I understand it, schema S is such that the cause properties are such that they could be had by either cause without their losing their identity as the cause that they are. Thus, in genuine instances of S, C1 could have had CP2 (rather than CP1) whilst retaining its identity as C1, or, in other words, without becoming C2. But this is not the case with the example under consideration for if C1 had had CP2 rather than CP1 (that is had it had the property of being C2 rather than that of being C1), it would not have been C1. Therefore, though it may well be the case that the property of being C1 is not a causal power, this example is not relevant to the issue in question; it would have been relevant only if it were an instance of S.

Fodor's second example features our old friends the property of being an H-particle and that of being a T-particle. The difference between Fodor's dime's being heads up and its being tails up is responsible for the difference between every particle in the universe being an H-particle and every particle in the universe being a T-particle. This, he thinks, is another non-bona fide instance of S, for 'the difference between being heads up and being tails up does not count as a causal power in virtue of its responsibility for this difference in the particles' (p. 11).

Again I doubt that this case is a genuine instance of S, and so question its relevance to the issue in question. For it to be a genuine instance of S there would have to exist a causal connection between Fodor's dime and every physical particle in the universe, for C1 and C2 in this case are Fodor's coin (or the event of its being flipped), and E1 and E2 are changes in each and every particle in the universe. In *Psychosemantics* Fodor seemed to rule out the possibility of such a causal relationship when he said:

how on earth could the causal powers of particles on Alpha Centauri depend on the orientation of my dime? Either there would have to be a mechanism to mediate this dependency, or it would have to be mediated by a fundamental law of nature; and there aren't any such mechanisms and there aren't any such laws. (p. 39).

If there are no causal mechanisms mediating the connection between Fodor's dime and the causal powers of distant particles, then presumably there are no causal mechanisms mediating causal connections between Fodor's dime and such distant particles.

Here is the third example. Fodor has siblings but his twin does not. In virtue of his having siblings Fodor is able to produce sons who are nephews, something that his twin cannot do. But, argues Fodor, it is not the case that the difference between having siblings and not having them is a difference in causal powers in virtue of its responsibility for the difference between having children who are nephews and having children who are not nephews.

I do not dispute that this example fits schema S, but I do not share Fodor's intuition that it is not a bona fide instance of schema S. As far as I can see, in being able to father nephews, Fodor has a power that his twin does not have. Of course that is not to say that this is a power that science should care about. No-one wants to argue that for every causal power there is some science that cares about it. Edgar has the power to pacify Fang, a power that no-one else has, but I take it that there is no science that assigns Edgar, and Edgar only, to a particular type in virtue of his having this power.

Perhaps there is not much profit to be had from merely refusing to endorse Fodor's intuitions concerning which instances of S are bona fide and which are not, for the danger is that the debate reduces into a squabble over intuitions. What is needed is a more substantial objection to his argument, an objection the construction of which necessitates the adoption of a different line of approach.

Condition C is such that '[o]nly when it is not a conceptual truth that causes that which differ in that one has CP1 where the other has CP2 have effects that differ in that one has EP1 where the other has EP2' (p. 19) is an instance of S a bona fide case of a difference in causal powers. The siblings example fails this test 'because having siblings is *conceptually* connected to having sons who are nephews: to be a nephew *just is* to be a son whose parents have siblings' (p. 19). Similarly, the H-particle case fails the test 'because the connection between all the world's particle becoming H-particles at time t and my coin's being heads up at T is conceptual. To be an H-particle at t *just is* to be a particle at a time when my coin is heads up' (p. 19).



An example of a case that satisfies C is the following. The difference between being a planet and not being one is a difference in causal powers in virtue of its responsibility for the difference between having a Keplerian orbit and not having such an orbit. That is so 'because it is true and contingent that, if you have molecularly identical chunks of rock, one which is a planet and the other which is not, then, *ceteris paribus*, the one which is a planet will have a Keplerian orbit, and *ceteris paribus*, the one which is not a planet will not' (p. 19).

Is condition C really necessary? In actual fact it does not comply with all of Fodor's intuitions as we shall now see. Fodor believes that for any contingent property it is nomologically possible to build a machine that reliably detects that property. An example of such a contingent property is that of having had a Bulgarian Grandmother. It is possible to construct a Bulgarian Grandmother detector, 'a machine which exhaustively examines the piece of space-time that starts with the birth of your Grandmother and ends with your birth and which goes into one state if it detects somebody who was your Grandmother and was Bulgarian . . . but which goes into another state in case it detects no such property' (p. 13).

Fodor asserts that 'having a Bulgarian Grandmother is having a causal power in virtue of the (actual or possible) effects that instantiations of this property have on Bulgarian G . . . detectors' (p. 14). Call the state the Bulgarian Grandmother detector goes into if it detects a Bulgarian Grandmother the positive state, and the state it goes into if it detects no such person the negative state. Then we can generate the following instance of schema S. C1 is Edgar and CP1 the property of having had a Bulgarian Grandmother. C2 is twin Edgar and CP2 the property of not having had a Bulgarian Grandmother. E1 and E2 are effects (actual or possible) that Edgar and his twin, respectively, have on Bulgarian Grandmother detectors. EP1 is the property of going into the positive state and EP2 is the property of going into the negative state. In this case the difference between having CP1 and having CP2 is a difference in causal powers in virtue of its being responsible for the difference between E1 and E2, i.e. the difference between having EP1 and having EP2. But this case does not satisfy condition C, for there is a conceptual relationship between the difference between having CP1 and having CP2, and the difference



between E1 and E2, that it is responsible for. Why? It is a conceptual truth that individuals who differ in that one had a Bulgarian Grandmother whilst the other did not, have effects on Bulgarian Grandmother detectors that differ in that one has the property of going into the positive state whilst the other has the property of going into the negative state. If you have the property of having had a Bulgarian Grandmother and you cause a putative Bulgarian Grandmother detector to go into a negative state, then the machine isn't really a Bulgarian Grandmother detector.<sup>85</sup> And if you cause a genuine, working, Bulgarian Grandmother detector to go into the negative state then you can't have had a Bulgarian Grandmother.

Fodor argued that the sibling case failed to satisfy condition C 'because having siblings is conceptually connected to having sons who are nephews; to be a nephew [EP] *just is* to be a son [E] whose parents [C] have siblings [CP]' (p. 19). If that is right then surely having had a Bulgarian Grandmother is conceptually connected to causing Bulgarian Grandmother machines to go into the positive state. This is so because to be a change in the state of a Bulgarian Grandmother detector into the positive [EP] *just is* to be a change in the state of a Bulgarian Grandmother detector [E] that is caused by exposure to individuals [C] who have the property of having had a Bulgarian Grandmother [CP]. Therefore the case fails to satisfy condition C, and so that condition doesn't sit happily with Fodor's intuitions after all. Moreover, given that Fodor would appear to be right in thinking of this case as a bona fide instance of schema S, it would appear that we have a reason for rejecting condition C.

What about the case of the Putnamian twins and the two putative counterexamples to Fodor's supervenience claim that they generate? Fodor believes that they fail to satisfy condition C. Consider the second case, that where a difference in the twins' causal connections to water is responsible for a difference in the intentional properties of their thoughts. He thinks that there is a conceptual connection between the difference between Oscar and Oscar2, and the difference between the broad intentional properties of their thoughts. He says 'it is conceptually necessary that if you are connected to water in the right way then you have water thoughts (rather than twater

---

<sup>85</sup> Or a *ceteris paribus* clause as been violated as would be the case if the machine was malfunctioning.

thoughts) (p.20), and that 'to have a water thought *just is* to have a thought that is connected to water in the right way' (p.21). No doubt this is right, but that is because "right way" is elliptical for "in the way that makes your thoughts water thoughts". In other words, it is no doubt true that "it is conceptually necessary that if you are connected to water in the way that makes your thoughts water thoughts then you have water thoughts (rather than twater thoughts)". But this fact need not worry the externalist, for it doesn't follow from it that there isn't some property that Oscar has (and that his twin doesn't have) that is responsible for his thoughts being water thoughts and that isn't conceptually connected to that effect property.

There are many different causal connections that an individual might bear to water. Not all of them are capable of supporting water thoughts. Call the specific connection that Oscar bears to water, a connection in virtue of which his water thoughts are water thoughts, CR1.<sup>86</sup> The question is: is the connection between bearing CR1 to water and having water thoughts conceptual? In other words is it conceptually necessary that if you stand in CR1 to water then you have water thoughts, or that to have a water thought *just is* to have a thought that is connected to water in relation CR1? My intuition is that we should supply a negative answer to this question, that causal relationships other than CR1 are capable of supporting water thoughts, or that if one must bear CR1 to water to have water thoughts the notion of necessity involved is more nomological than conceptual. I'm not going to attempt to vindicate this intuition but rather argue that if Fodor rejects it he runs the risk of contradicting other central aspects of his thought. Before I do this I will consider Fodor's response to the first Putnamian case which, naturally enough, is very similar to the one we have just considered and is open to the same kind of objections.

In this case the difference between the twin's thoughts (one has water thoughts whilst the other has twater thoughts) is responsible for a difference in the broad intentional properties of their behaviour (namely between Oscar's behaviour's being water behaviour and

---

<sup>86</sup> CR1 is to be distinguished from, inter alia, CR2, the relationship that Oscar2 bears to water just after he has been transplanted to Earth. When Oscar2 stands in CR2 to water, he bears a causal connection to water, but not a causal connection that is such as to make any of his thoughts water thoughts.

Oscar's behaviour being water behaviour). Fodor argues that this case fails to satisfy condition C, for the difference between the causes is conceptually connected to the difference between the effects. This is so as 'it is conceptually necessary that people who have water thoughts (rather than twater thoughts) produce water behaviour. . . Being water behaviour *just is* being behaviour that is caused by water thoughts (rather than twater thoughts)' (p.21). No doubt this is all correct, but so what? Oscar's water thoughts have the property of being the thoughts of someone who bears CR1 to water, a property the having of which is responsible for his water thoughts having, as their behavioural effects, water behaviour. If Oscar stood in CR2 to water, then none of his behaviour would be water behaviour. Here we have an example of schema S that doesn't obviously fail condition C, for it is not obvious that there is a conceptual connection between the relevant property of Oscar's water thoughts (i.e. their being the thoughts of someone who bears CR1 to water) and their behavioural effects being water behaviour. It is not obvious that being water behaviour *just is* being behaviour that is caused by the thoughts of someone who bears CR1 to water.

Of course none of this conclusively establishes that the Putnamian twins satisfy condition C, but it does dent Fodor's argument somewhat. Moreover, as I said above, to reject my intuition that, for example, there is no conceptual connection between bearing CR1 to water and having water thoughts is perhaps not an option open to Fodor for I think it conflicts with other aspects of his overall view.

This conflict has to do with the theory of content. Over recent years one of Fodor's main preoccupations has been to construct a theory of content, a theory that explains, in naturalistic terms, why our thoughts have the semantic properties that they have. The theory he has developed is a descendent of the informational semantics of Dretske (1981) and Stampe (1977), and takes the form of the specification of a sufficient condition for a mental representation's expressing the property that it expresses.<sup>87</sup> What makes Oscar's water thoughts water thoughts is that he (or those thoughts) bear

---

<sup>87</sup> Interestingly, in the present connection, Fodor doesn't think he has a necessary condition for a mental representation's expressing the property that it expresses, but only a sufficient condition, a sufficient condition that our mental representations satisfy.

relation CR1 to water. Suppose there exists a conceptual connection between bearing this causal relation to water and being a water thought. Then the story of why Oscar's water thoughts are water thoughts will be conceptually true just as the story concerning why I am a nephew is conceptually true (because I have a parent who has a sibling). Moreover, the question of why his water thoughts are water thoughts will be answerable on the basis of an a priori reflection on the concepts involved (just as the question of why I am a nephew can be answered on the basis of a priori reflection). And, of course, the same will be true of everyone else's water thoughts so that a general answer to the question "what makes a water thought about what it is about?" can be answered solely on the basis of a priori reflection and that answer will be conceptually true.

Now, presumably, there is nothing special about water thoughts. Dog thoughts, for example, will be about dogs in virtue of the causal relationship that they bear to dogs, and similarly for any thought about any object or property instantiated in the external world that any individual is capable of having. Therefore, for any type of thought there will be a general answer to the question "what makes thoughts of that type about what they are about?", an answer that can be arrived at solely on the basis of a priori reflection, and an answer that will be conceptually true.

Presumably the causal relationship that one's water thoughts must bear to water to be water thoughts will mirror the causal relationship that one's dog thoughts must bear to dogs to be dog thoughts. And, quite generally, for any X, the causal relationship that one's X thoughts must bear to X to be X thoughts will mirror the causal relationship that Oscar's water thoughts bear to water in virtue of which they are water thoughts. In short, the story about Oscar's water thoughts will be a particular instance of a general story about why thoughts are about what they are about. The upshot of all this is that a general theory of content can be arrived at by a priori reflection, a theory that will be conceptually true.

I do not believe that Fodor would be happy to endorse this upshot.<sup>88</sup> The view that there is a conceptually true theory of content to be had,

---

<sup>88</sup> However, having said that, his approach in constructing a naturalistic theory of content does have a distinct a priori air to it (see Chapter 7 for the details). But even if Fodor would be happy to endorse the consequence that I am describing, the point



a theory obtainable by a priori means, is not the sort of view that one expects Fodor to hold. His holding such a view would be in tension with his general views on the nature of philosophy and on how the philosopher should proceed. Fodor has championed the view according to which there is no real, clean-cut distinction between philosophy and science. One cannot get very far, he thinks, by a priori reflection alone, hence scientific investigations and findings are relevant to philosophical concerns. Thus the various theories that Fodor has championed during his career<sup>89</sup> were never held by him to be conceptually or necessarily true, or provable by a priori means. Thus it would be surprising, if not contradictory, were Fodor to endorse the upshot of rejecting my intuition that there is no conceptual connection between bearing CR1 to water (where CR1 is the relation that Oscar/his water thoughts bear to water in virtue of which his water thoughts are water thoughts) and having water thoughts. But if he doesn't reject this intuition it would appear that he has to accept that the putative Putnamian counterexamples are genuine counterexamples to his supervenience thesis.

In the spirit of the point made in the preceding paragraph, it might reasonably be thought that there is tension between Fodor's line in 'A Modal Argument for Narrow Content' and another view that he holds dear, and which plays a fundamental role in his overall system of thought. Functionalism as a theory of content is the doctrine that the content of a mental state is determined by its causal relations to other mental states. The idea is that, for example, my thought that there is a cat on the mat has the content *there is a cat on the mat* in virtue of what other thoughts it is disposed to cause/ I am disposed to infer from it. So that, for example, it wouldn't have that content if I wasn't disposed to infer from it a thought with the content *there is an animal on the mat*. On the face of it this idea is quite plausible and sits happily with the intuition that in order to have cat thoughts one has to believe a whole load of true things about cats, one belief amongst these being that cats are animals.

---

remains that he should not be so happy, for he cannot consistently hold that there is a conceptually true theory of content to be had, a theory that is obtainable by a priori means.

<sup>89</sup> For example, the Representational Theory of Mind and methodological solipsism.



Now Fodor believes that there is no analytic/synthetic distinction: 'I take it very seriously that there is no principled distinction between matters of meaning and matters of fact. Quine was right; you can't have an analytic/synthetic distinction' (1990, p. x). The absence of such a distinction has a quite a dramatic consequence if functionalism is true, this being 'that you can't have a principled distinction between the kinds of causal relations among mental states that determine content and the kinds of causal relations among mental states that don't' (p. x).<sup>90</sup> Barring an accident on a cosmic scale, take any thought of mine you like and no-one else will have a thought with just the same causal relations to other mental states. Therefore no-one else will have a thought with just the same content and, as thoughts have their contents essentially, no-one else will ever share a thought with me. The same will go for all other folks, so that, barring an accident on a cosmic scale, no two individuals will ever share a thought. Consequently intentional laws will subsume one individual at most, which means that there are effectively no intentional laws, and thus no scientific intentional psychology. This consequence of the functionalist theory of content - given an absence of any analytic/synthetic distinction - motivates Fodor's attempt to develop an atomist alternative.

Now we are in a position to notice the contradiction which, quite simply, is this: how can someone who avowedly rejects the analytic/synthetic distinction (and whose rejection of which plays such an important role in his thought) help himself to a notion of conceptual connection? If 'there is no principled distinction between matters of meaning and matters of fact', how can there be any principled distinction between conceptual and non-conceptual connections. And if there is no such distinction how can Fodor legitimately appeal to the notion of a conceptual connection to defend his supervenience claim?

---

<sup>90</sup> If there were an analytic/synthetic distinction one could claim that only certain privileged causal connections determined the content of a thought. Thus, for example, in order for one to have cat thoughts one must believe that cats are animals, but one need not believe that ocelots are members of the cat family. Hence I need not be disposed to infer from my belief that there is a cat on the mat a belief that there is something related to an Ocelot on the mat in order for the former belief to be the belief that there is a cat on the mat.

It might be thought that regardless of whether Fodor can legitimately appeal to conceptual connections, he is on to something in arguing in the manner in which he does. The idea I have in mind is that scientific psychology must produce explanations, specify connections, and appeal to generalisations that are contingently rather than conceptually true. Consequently, the property of being a water thought and that of being water behaviour cannot legitimately figure in the taxonomy of scientific psychology, as explanations and generalisations that appeal to them will be conceptually rather than contingently true. I think that there is a way of answering this line of thought, and doing so in such a way as to reconcile the Humean intuitions that underlie it with the idea that scientific psychology should recognise the difference between water thoughts and twater thoughts. Here goes. There may well be a conceptual connection between water thoughts and water behaviour (in the sense that water behaviour *just is* behaviour that is caused by water thoughts) and it may well be conceptually true that my water thoughts cause water behaviour rather than twater behaviour. But that notwithstanding, the specific causal explanations that an externalist psychology would produce are contingent as would be the generalisations subsuming specific water thoughts. To see this consider the following. Suppose I switch on an empty kettle. Why did I do that? Because I wanted to boil some water and believed the kettle was full of water. Surely that is a legitimate causal explanation. The connection between my water thoughts and my behaviour is contingent as is the explanation of my behaviour. If you want to know why I am a nephew you can generate from your armchair the following conceptually true explanation: because I have a parent with siblings. But the explanation of why I switched the empty kettle on is nothing like that. My behaviour could have had just about any cause, and which cause it had (and thus how it is to be explained) is a matter for empirical investigation. Moreover, I could have had just those thoughts without switching the kettle on, in the respect that there are conceptually possible worlds where those very thoughts have quite different causal powers.<sup>91</sup> In those possible worlds the generalisations that subsume

---

<sup>91</sup> To see that there is nothing conceptually necessary about the specific causal powers of the various water thoughts that we are capable of having, consider the

our water thoughts diverge from those that hold in the real world. Therefore, one can assert that a scientific psychology can appeal to specific water thoughts in order to explain specific mental and behavioural episodes, and can employ generalisations in which reference to such water thoughts figures, without abandoning one's Humean intuitions. Hence an appeal to Humean considerations gives no grounds for the conclusion that scientific psychology shouldn't distinguish between water thoughts and twater thoughts.<sup>92</sup>

I thus conclude that Fodor's modal argument fails to establish that the Putnam case doesn't constitute a genuine counterexample to the claim that the causal powers of psychological states are locally supervenient. By way of recapitulation, here is how the argument went. Fodor argued that the difference between being a water thought and being a twater thought is not a difference in causal powers in virtue of being responsible for the difference between its behavioural

---

following example. Suppose that walking through a forest I come across a pool of deep water and form the belief that there is a pool of deep water before me. This belief will interact with other thoughts of mine to produce certain mental and behavioural effects. We can imagine a creature capable of believing that there was a pool of deep water before it such that whenever it had that belief it went into a state of fear (perhaps it belonged to a species that had evolved in an environment where the deep pools of water tended to be full of crocodiles or other forces of destruction). Thus, in this creature the belief in question has causal powers that it doesn't have in me, for to cause fear in me it has to interact with other beliefs that I am not guaranteed to have (in the imaginary creature the belief automatically causes the state of fear without having to interact with any other beliefs). In other words, which causal powers my belief that there is a pool of deep water before me has is a contingent matter, in that it is conceptually possible that a creature could have had just the same belief yet that belief had different causal powers. I take it that the point generalises to all water thoughts.

<sup>92</sup> If Fodor denies this claim about the contingency of the causal connections, explanations, and powers of our specific water thoughts then it is difficult to see how he can avoid concluding that the narrow psychology that he envisages doesn't fall foul of just the same Humean considerations. For, given the close relationship between narrow and broad content, if it is conceptually true that my belief that there is water in the kettle and my desire to boil some water causes me to switch on the empty kettle, then it will be conceptually true that my narrow belief and my narrow desire causes the narrow behaviour that I engage in.

effects being water behaviour and twater behaviour. This is because this case fails to satisfy a necessary condition for being a difference in causal powers. In response, I argued that: (i) this putative necessary condition does not cohere with Fodor's intuitions and thus is robbed of much of its motivation; (ii) that the Putnam case may well pass Fodor's test for being a difference in causal powers; (iii) that were Fodor to reject my argument for (ii) he would run the risk of inconsistency; (iv) that Fodor's argument clashes with other important aspects of his overall view; and (v) that it is not in conflict with reasonable Humean considerations to allot water thoughts to different scientific psychological types than their corresponding twater thoughts.

## 5.6 The nature of causal powers

None of the foregoing establishes that a psychology that individuates in terms of causal powers will not respect the supervenience of the psychological on the physical; rather, all it does is undermine Fodor's attempt to establish that the Putnamian case doesn't constitute a bona fide counterexample to his individualism. Hence, the question of whether scientific psychology must respect local supervenience is still very much unsettled. Given this, it would perhaps be instructive to examine the notion of a causal power in somewhat more detail than Fodor does, and then attempt to apply any results of this examination to the case of psychology.

For an entity, state, or event to have a given casual power, it must be capable of causing an effect that constitutes an exercise of the power in question. Such an effect need not, in actual fact, be caused. For example, for me to have the power to lift Fang I must be capable of causing an event of my lifting Fang but it is not necessary that I have ever lifted Fang. It is enough that that effect is produced in some relevant (possibly counterfactual) context. Thus, as Fodor tells us, causal powers are to be assessed across contexts; when one is addressing the question of whether a thing has a given causal power one must determine not just its actual effects but the effects it produces in a range of counterfactual contexts. However, this raises the question of which contexts are relevant. There are reasons for resisting the conclusion that all contexts are relevant, as can be seen



from reflecting on the property of being a planet. Fodor holds that being a planet does affect causal powers, for being a planet affects your trajectory, which in turn affects what you can bump into. This thought seems intuitively right. Now a planet could have a molecularly identical twin that was not a planet. For this non-planet not to have the same causal powers as its planet twin, contexts in which it has the relational property of orbiting a star must not be relevant to its causal powers. If such contexts are relevant then the property of being a planet would not affect the causal powers of whatever had it, and thus would not figure in the taxonomy of astronomy. So the question arises of which contexts are relevant when assessing causal powers, and in virtue of what are the relevant contexts relevant? In order to answer these questions, and to shed more light on the notion of a causal power in general, it will be instructive to examine an example of a legal power.

Edgar and Waldo both rent rooms in a house owned by a Mr. Higgins. Edgar finds Waldo a pain to live with due to the latter's continually playing loud music, keeping late hours, and poor standards of personal hygiene. Edgar wishes he could evict Waldo but does not have the legal power to do so. Higgins does have the legal power to evict Waldo, but, despite Edgar's exhortations, he refuses to exercise this power. ("So long as he pays the rent what do I care?" is what he says). So Higgins has a legal power that Edgar does not. In virtue of what do they differ in this respect? They clearly do not differ because of the effects that they produce, for Higgins does no evicting. In line with the thought that legal powers are to be assessed across contexts, the obvious answer is that what is crucial is that there are counterfactual contexts where Higgins produces the effect of evicting Waldo but no counterfactual contexts in which Edgar produces such an effect. This answer is on the right lines, but there is a problem with it, as there clearly are counterfactual contexts where Edgar evicts Waldo (contexts in which, for example the laws of the land diverge from those that hold in reality, or where Edgar is Waldo's landlord). Thus in order to appeal to differences in effects produced in counterfactual contexts when accounting for the difference between the legal powers of Edgar and Higgins, we must find a way of ruling out such contexts as irrelevant. How are we to do this?



When evaluating the legal powers of an individual, not all possible contexts are relevant. When we ask whether an individual has a given legal power LP, what we are asking is whether they, largely as they are now, in the world, largely as it is now, are capable of producing an effect that constitutes an exercise of LP. Hence the relevant contexts are those where the individual and the world is as it is now in all relevant respects. But what are these relevant respects? It is difficult to frame a general answer to this question, but in connection with particular cases it is easy enough to appreciate what is relevant and what is not. In the present case contexts in which the laws of the land differ from those that hold in reality (particularly those relating to the rights and obligations of landlords and tenants) are not relevant. For example, that Edgar succeeds in evicting Waldo in a counterfactual context where it is enshrined in law that you can evict a fellow tenant if he goes six months without cleaning the bathroom clearly doesn't imply that he has the power to evict Waldo. Similarly, contexts in which Edgar stands in relation to Waldo and Higgins differently than he does in the real world are not relevant. For example, that Edgar evicts Waldo in a context where he is Waldo's landlord doesn't imply that he has the legal power in question in the actual world. What this suggests is that when assessing an individual's legal powers we must hold fixed the laws of the land and those properties of his (in particular his relations to his fellows) which determine how the laws of the land impinge upon and apply to him.

Similar considerations apply to the case of causal powers. Thus whether or not something has a causal power P depends not just on the effects that it actually produces but the effects that it would produce in certain counterfactual contexts. But not all counterfactual contexts are relevant. When we ask whether Edgar has the power to lift Fang, we are asking whether Edgar, as he is now, in the world as it is now, has the ability to lift Fang as he is now. Contexts where Edgar is stronger as a result of being a weight lifter, where Fang is heavier, and where the laws of nature are other than they are in the real world, are not relevant contexts. Thus when we are investigating whether or not Edgar has the power to lift Fang we must hold fixed such properties of Edgar, Fang, and the world in which they live considering only what happens in such contexts.

Can we generalise this point? Intuitively, what we want to say is that when assessing causal powers we must hold fixed all those properties of the item under consideration and its world that determine whether or not it has the causal power in question. To put the matter in these terms runs the risk of circularity; what we need - or so it would appear - is a non-circular characterisation of which counterfactual contexts are relevant and thus should be examined when evaluating causal powers. It might be thought that we could get round this problem by saying that all we need to hold fixed are the laws of nature that hold in the actual world. However this won't do, for in the case of the question of whether Edgar has the power to lift Fang, the contexts where Fang is a different weight than he is in real life, or where Edgar's physical strength is different than it actually is are both nomologically possible but clearly irrelevant. Thus, besides laws of nature, we must hold fixed both properties of the item under consideration and its environment when assessing causal powers. The properties that we must hold fixed are those which, given the laws of nature, determine whether or not the item under consideration has the causal power in question.<sup>93</sup>

I don't think I can formulate a general, non-circular account of what properties of an item one must hold fixed when assessing its causal powers. But this is not really a disaster as enough light has been shed on the notion of a causal power to enable some progress to be made on the discussion of the question of whether a scientific psychology that individuates in terms of causal powers must respect the supervenience of the psychological on the physical.

---

<sup>93</sup> To bring out the similarity between the case of legal powers and that of causal powers, consider the following: The laws of nature in the case of causal powers are the analogue of the laws of the land in the legal powers example. To not hold them fixed when assessing causal powers would be as absurd as to argue that Edgar has the power to evict Waldo on the grounds that he evicts Waldo in a counterfactual context where it's a law that you can evict your fellow tenants if they don't clean the bathroom. The properties of the entity under consideration that we must hold fixed when assessing causal powers are the analogues of such properties as that of Edgar's being a tenant, his bearing such and such relations to Waldo and Higgins, and so on. To not hold such properties fixed when assessing causal powers would be as absurd as to conclude that Edgar has the power to evict Waldo on the grounds that he evicts Waldo in a counterfactual context where he is Waldo's landlord.

I have argued that there is a similarity between causal powers and legal powers, but there is a characteristic of the latter that it offends intuition to ascribe to the former. Legal powers are not locally supervenient; that is, two individuals can be molecule for molecule identical without having the same legal powers. However, there is a very strong intuition that if two entities are molecule for molecule identical then they must agree in their causal powers given that causal powers are to be assessed across contexts. For, the thought goes, for any context you care to consider, if one of the twins causes an effect in that context, its duplicate would produce just the same effect were it slipped in its place. Perhaps this is more than just an intuition but, rather, an idea that is intimately bound up with the physicalist view that our world is at bottom a physical system populated by physical entities that are governed by the laws of physics. If you are a physicalist then it is very tempting to suppose that the events that a system produces in a given context will be determined by its physical properties so that a physical duplicate would produce just the same effects were it placed in that very context. From this it's a short step to the conclusion that physically identical entities will agree in their causal powers. I suspect that this kind of intuition is at work in Fodor's reflections.

On the other hand, it does seem possible for physically identical entities to diverge in their causal powers. Thus, for example, surely a planet would have different causal powers than a molecularly identical non-planet that was no more than a stationary chunk of rock. Another example would be that of a meteor. Two rocks could be physically identical but one be a meteor and the other not, the latter being a stationary rock. Surely the meteor, hurtling through space as it is, will have the power to cause craters, a power that its stationary twin lacks. For how can a rock that just sits there motionless have the power to cause any craters?

It would appear that we have a case of conflicting intuitions. However, the conflict is only *prima facie*, for there is a way of resolving it that does justice both to the thought that physical twins have identical causal powers *and* to the thought that such properties as that of being a planet and that of being a meteor affect the causal powers of whatever has them. The way in which the putative conflict resolves has bearings on the issue of whether a psychology

that individuates in terms of causal powers must respect supervenience.

Consider the case of two molecularly identical rocks, one of which is a meteor, and the other of which is not. The meteor has the power to cause craters; to see this just reflect upon what happens when a meteor, hurtling through space as meteors do, crashes into a planet. Does the non-meteor have the power to cause craters? Well, in the context where it is hurtling through space and crashes into a planet it causes a crater. Whether this means that it has the power to cause craters all depends on whether this context is one that is relevant to assessing the non-meteor's causal powers. Notice that in this context the non-meteor will be a meteor and so the question of whether the context is relevant turns into the question of whether the non-meteor's property of being a non-meteor is one of those properties that is to be held fixed when assessing its causal powers. Now for the crucial point: whether this property is to be held constant depends upon how one sees or considers the non-meteor, and how one so sees or considers it is an interest relative matter. Considered astronomically, it is the astronomical properties of the rock that are to be held fixed when assessing its causal powers. As far as the astronomer is concerned, there is all the difference in the world between being a meteor and not being one, hence the context where our non-meteor is hurtling through space is one where its astronomical properties differ considerably from those that it has in the real world. Therefore this context is not relevant when assessing the non-meteor's causal powers so long as that non-meteor is considered astronomically. The matter could be put this way: *qua* astronomical body, meteors have a causal power, namely the power to cause craters, that their stationary non-meteor twins do not have. In this respect the property of being a meteor affects the causal powers of whatever has it.

However, one is not forced to consider the meteor and its non-meteor twin astronomically, for one can consider them as physical bodies. And *qua* physical bodies they have just the same causal powers, or at least they both have the power to cause craters. Here's why. The properties of the rocks that the physicist will hold fixed when determining their causal powers will be those properties that are instances of properties that belong to the set of properties that



physics cares about. The property of being a meteor and that of being a non-meteor do not belong to this category and so they are not to be held fixed. Therefore the context where the non-meteor is hurtling through space (and thus is a meteor) is a relevant context when determining whether it has the power to cause craters. The properties of the rocks that are to be held fixed are their internal physical properties; thus the relevant contexts are all those nomologically possible contexts where the rocks are, in terms of their internal physical constitution, just as they are in the actual world. Given their physical identity, the rocks would produce identical physical effects in each and every nomologically possible context and so, qua physical bodies, they have just the same causal powers. Thus what resolves the above *prima facie* conflict is the fact that qua physical bodies the rocks have just the same causal powers (in particular they both have the power to cause craters) but qua astronomical bodies they do not (for the meteor has the power to cause craters, a power that the non-meteor does not have).

In short, then, when assessing an entity's causal powers, its effects across a range of contexts are to be considered. What all these contexts have in common is that they are nomologically possible; the laws that hold in them are just the same as those that hold in the actual world. Yet it is often the case that not all nomologically possible contexts are relevant, for when assessing causal powers one must hold fixed certain of the properties of the individual in question (and sometimes, also, properties of its environment). But which of these properties are to be held fixed is an interest relative matter; it all depends upon what science you are doing and how you consequently see the entity in question.

What bearing does all this have on the question of whether psychology individuates in such a way as to respect the supervenience of the psychological on the physical? The astronomical example suggests that from the point of view of the special sciences, causal powers are not always locally supervenient. This should make us alive to the possibility of such a violation of supervenience in the case of psychology. Perhaps there could be the following kind of case. Twin1 and Twin2 are a pair of individuals who are molecule for molecule identical. Twin1 has a psychological state, namely PS1, that stands in a relation of identity to (or is



constituted by) one of his brain states. The corresponding psychological state of Twin2, PS2, stands in a relation of identity to (or is constituted by) one of his brain states. Qua physical state, PS1 and PS2 will have the same causal powers, and it may well be the case that qua brain state they have the same causal powers. But it doesn't follow from this that qua psychological state they agree in their causal powers. It doesn't follow because these states may differ in some respect that the psychologist cares about - a respect that neither the neurophysiologist or the physicist cares about - that makes a crucial difference to which contexts are relevant when assessing their causal powers. The kind of psychological difference that I have in mind narrows down the contexts relevant in determining the causal powers of PS1 and PS2 (qua psychological states) in such a way that they turn out to have different causal powers. For example, suppose PS1 has a property P1 that psychology cares about but which PS2 does not have. (Such a property would be analogous to the astronomer's property of being a meteor). In assessing PS1's causal powers, only those contexts in which it has P1 are to be considered. In some of those contexts PS1 produces a psychological effect which counts as an exercise of a psychologically significant causal power CP. In virtue of this, PS1 has the causal power CP. PS2 only produces effects that count as an exercise of CP in counterfactual contexts where it has the property P1. But these contexts are not relevant given that PS2 in reality doesn't have the property P1. Therefore PS2, qua psychological state, doesn't have the causal power CP, and as CP is a psychologically relevant causal power, PS1 and PS2 belong to different psychological types. The upshot of this is that the twins, despite their being physically identical (and perhaps neurophysiologically identical) are not psychologically identical.

Of course this doesn't establish the conclusion that psychology, even though it individuates in terms of causal powers, does not respect the supervenience of the psychological on the physical. To do that, we would need to produce some concrete example or engage in an examination of the explanatory ambitions and assumptions that are specific to psychology.<sup>94</sup> Moreover, it might be objected that there is a

---

<sup>94</sup> I do not wish to argue that the Putnam case constitutes such an example. The externalist is not, solely in virtue of his externalism, committed to the thesis that

fundamental feature of the planet and meteor case that rules out the possibility of a psychological analogue. The meteor has different causal powers than its non-meteor twin, but that is perfectly explicable given the fact that the two chunks of matter are subject to different proximal stimuli. There are forces acting on the surface of the meteor that push and pull it around its local environment in a way quite different to those acting on the surface of its twin. (Forces such as gravitational pull, air pressure, and the like). Hence, it is not surprising that the meteor has the power to cause craters; its having that power is a direct product of the proximal stimuli it is subject to. Yet, so the objection continues, in the kinds of cases the externalist dreams up, the proximal stimuli the imagined twins are subject to are intrinsically identical, so making it inexplicable that they could have different causal powers. In short, Fodor can quite happily concede that causal powers are not always locally supervenient whilst holding onto the idea that the causal powers of psychological states are locally supervenient.

My first response to this objection involves pointing out that there are many cases of differences in causal powers in the absence of differences in proximal stimuli. Consider a particular Royal Bank of Scotland cash dispenser. It has the causal power to dispense cash and to effect changes in the bank accounts of individuals to whom it dispenses cash. It is easy to imagine a twin of this machine that inhabited an environment where the pieces of paper that it dispensed were not bank notes but concert tickets. This second machine does not have the power to dispense cash, despite the fact that it might be subject to just the same proximal stimuli as the first machine. That there is such a difference in causal powers (and a significant one at that) is perfectly explicable: it is due to differences between the respective environments of the machines, differences that are underpinned by physical differences at some distance from the machines. In this case, the context where the cash dispenser is transported to the environment of its twin so that it no longer dispenses cash is not relevant when assessing its actual causal powers.

---

Oscar and Oscar2 are psychologically divergent. The widespread discussion of the Putnam case makes it easy to forget that establishing that the twins are psychologically identical does not thereby defeat externalism.

My second response involves trying to construct a concrete psychological example. The example I have in mind played a prominent role in my reflections in Chapter 3. I have the cognitive capacity to recognise certain individuals. I exercise this capacity when I form the true belief that Fang is before me whenever I am confronted by Fang. Underlying and facilitating this capacity are subpersonal representational states that carry information about specific individuals. That they carry information about specific individuals, that they represent those individuals as having certain properties, is crucial to their facilitating the exercise of the cognitive capacity in question. Consider the representational state in me that represents properties of Fang. It has the power to enable me to recognise Fang. The corresponding representation in my twin represents some other dog, and thus does not have the power to enable him to recognise Fang (in fact, were he to see Fang he would come to the mistaken belief that some other dog was before him). Thus, our respective states have, qua psychological states, divergent causal powers, as the context where I am embedded in my twin's environment (a context where my state represents some other dog as having certain properties) is not relevant to assessing the causal powers of my state. And surely the causal power of enabling me to recognise a specific individual that I can in fact recognise is one that scientific psychology does and should care about. Therefore, my state and the corresponding state of my twin, in differing in the specific individual that they represent and carry information about, differ in a psychologically significant way.<sup>95</sup> So the very case that I earlier appealed to in order to justify the claim that scientific psychology does and should appeal to properties of psychological states that are not locally supervenient would appear to provide us with just the

---

<sup>95</sup> In Chapters 2, 3, and 4 I emphasised the role of contingent facts about the world in facilitating cognition. An upshot of this point is that it is contingent that my state (with the information that it carries and the content that it has) enables me to recognise Fang and enables me to recognise Fang in the way described by the scientific psychologist. Therefore, to describe this state as differing from that of my twin in a psychologically significant way is not to violate any Humean considerations. Whatever one thinks about the Putnam case, it looks as if we here have an instance of schema S that satisfies Fodor's necessary condition for being a difference in causal powers.

kind of example that we need to fend off Fodor's metaphysical argument for individualism.

Once again I accept that my considerations are hardly decisive; the individualist may well object that I have not described a plausible case of psychological states of twins that differ in their causal powers. In short, there is a real danger of a stand off. However, I think that I have done enough to blunt the power of Fodor's metaphysical argument by establishing that it is far from clear that it vindicates his individualist position. And this, allied with the considerations of Chapters 2 and 3, licenses the tentative conclusion that scientific psychology both does, and can quite legitimately, appeal to intentional properties that are not locally supervenient, and that it does not, nor need not, employ any notion of narrow content.

## 5.7 Conclusion

Before closing this chapter, here is a brief summary of the course that it has taken. Fodor thought that he had a metaphysical argument for the claim that scientific psychology does, or should, individuate in such a way as to respect the supervenience of the psychological on the physical, an argument that rests on the idea that sciences individuate in terms of causal powers. The case of the Putnamian twins poses an immediate problem for such a thesis for they, or some of their psychological states, would appear to diverge in their causal powers. Fodor provided two counter-arguments neither of which, or so I argued, are completely convincing. If we examine the notion of a causal power in greater detail than Fodor ever does it becomes clear that, from the point of view of special sciences, causal powers need not be locally supervenient. This raises the very real possibility that, in a manner akin to astronomy, scientific psychology does not individuate in such a way as to respect local supervenience, even if it does individuate in terms of causal powers. And that this possibility is more than a mere possibility is suggested by a consideration of some of the representational states that underlie and facilitate our recognitional capacities.

## Chapter 6

### The Naturalisation Project

#### 6.1 Introduction

One of the central preoccupations of contemporary philosophy of mind is that of constructing a naturalistic theory of content and evaluating various candidate theories. Reflecting this fact - perhaps even explaining it - the development of such a theory has been one of Fodor's primary concerns over the last decade or so.<sup>96</sup> Fodor has presented a theory inspired by the informational semantics of Dretske (1981) and Stampe (1977)<sup>97</sup>, central to which are the notions of mind-world causal laws and the relationship of asymmetric dependency between such laws. In the next chapter I will attempt to evaluate this theory. I will argue that, despite its impressive ingenuity and *prima facie* plausibility, it ultimately fails. However, before turning to an explicit consideration of Fodor's theory I will attempt to get clear on the nature of the problem that it attempts to solve; just what are those who seek to construct a naturalistic theory of mental content up to? It is especially important to do this given

---

<sup>96</sup> This marks a distinct change in Fodor's views for prior to the mid-eighties he seemed to hold that though mental content can in principle be naturalised (as would have to be the case for intentional psychology to be a respectable science) to attempt to produce a naturalistic theory of content would be to engage in an inevitably fruitless endeavour at this point in our intellectual history. For it is only when the rest of science has been completed that we will be in any position to explain in naturalistic and non-question-begging terms why our mental states have the intentional properties that they have.

<sup>97</sup> Fodor represents the informational theory as identifying the content of a symbol with the information that it carries, and as reducing information to causal covariation so that, for example, my *cow* thoughts are about cows because they are caused by, and only by, cows. This is, to put it mildly, a gross simplification of Dretske's complex and ingenious theory. However, we need not worry about this point in what follows.



the increasingly large band of philosophers who think that to attempt to naturalise mental content is to engage in a misguided project, a project that cannot, nor need not, be successfully accomplished and which is bound up with some disreputable scientific and reductionist assumptions. In this chapter I will attempt to present an account of the naturalisation project that reveals it to be important, sensible, and certainly not misguided. For if a scientific psychology of the kind that I have described in earlier chapters is to count as a respectable science then there is an important respect in which the semantic and intentional properties that figure in its theories and explanations must be naturalisable.

## **6.2 The place of mental content in the natural world**

The fundamental worry of the naturalist is that intentional mental phenomena have no real place in the natural world and that the intentional and semantic properties routinely ascribed to our mental states in the course of psychological description and explanation have no reality. Alleviating this worry would involve showing that, and how:

- (i) intentional mental phenomena occur in, and belong to, the natural world in broadly the same way in which, for example, biological, chemical, and geological phenomena do; and
- (ii) the intentional and semantic properties of mental phenomena are as real as, and as real in broadly the same kind of way as biological, chemical and geological properties.

The purpose of engaging in the naturalisation project is to alleviate this worry. This point could do with much in the way of expansion and clarification.

By definition scientific intentional psychology is committed to the idea that some of our mental states have intentional and semantic properties; more specifically that some of our mental states have semantic or intentional content. Moreover, its practitioners seek to produce explanations and theories in which appeals to such properties figure prominently. Consequently, if there are any requirements or conditions that a property must satisfy in order to be

scientifically respectable - in order to have a legitimate role to play in the theories and explanations of a respectable science - then they must be satisfied by the semantic and intentional properties that figure in psychological theory and explanation.

An assumption that unites those who engage in the naturalisation project is that there is just such a condition or requirement in virtue of the truth of the doctrine of physicalism and/or the commitment of science to that doctrine. Understood in the relevant sense, physicalism is the (perhaps vague) doctrine that our world is at bottom physical in nature; that it is at bottom nothing more than a very complex collection of physical particles organised hierarchically into aggregates that compose or constitute larger scale physical entities, states, events and processes, all of which behave in accord with the laws of physics. Thus physics is the most basic or fundamental science. However, according to the physicalist, this is not to say that physics is the only science or that physical properties are the only real properties. Rather, there are sciences other than physics - namely the special sciences - that appeal to properties that do not figure in the explanations, descriptions, predictions and laws of physics. According to the physicalist the phenomena that the special sciences study and the properties that they appeal to are related to physical properties and phenomena in roughly the following way. Each and every special science entity, state, event or process is possessed by something (be it an entity, state event or process) that is physically composed, constituted or realised. And the relationship between special science properties, on the one hand, and physical properties, on the other, is that the former either reduce to or supervene on the latter so that the totality of physical facts fixes or determines the totality of special science facts. In other words the world couldn't be different at the special science level without being different at the physical level.

The upshot of all this is that the physicalist is committed to the idea that for any putative special science to be in good repute it must satisfy the following condition:

The phenomena that fall within its domain of enquiry must be physically composed, constituted or realised; the properties that its laws and explanations quantify over must be possessed by phenomena that are physically composed, constituted or realised;

and its properties must reduce to, or supervene upon, the physical.

So the question arises: does intentional psychology satisfy the above condition? It might be thought that a positive answer can quickly be established by arguing that intentional mental phenomena are physically composed, constituted or realised via being neurophysiologically composed, constituted or realised and that intentional mental properties are possessed by phenomena that are at bottom clearly physical (be the bearers of such properties human individuals, brains, brain states, or something else along those lines). However this is far too swift, for it still remains to be shown that intentional properties either reduce to physical properties or supervene on the physical. A failure to indicate that (and how) lower level physical properties or facts fix, determine, or ground the intentional properties of mental states counts as a failure to indicate that (and how) intentional psychology manages to be a respectable, bona fide special science.<sup>98</sup>

In short then, to engage in the naturalisation project is to attempt to indicate that (and how) physical properties or facts fix, determine, or ground the semantic and intentional properties that are attributed to our mental states in the course of psychological theorising and explanation. And Fodor's theory of content is an exercise in that project.

Of course one might attempt to naturalise any putative special science property, for they must all be related to the physical in the above described manner if they are to be scientifically respectable. But there is a widespread feeling within the philosophical community that there is an element of added urgency in the case of the intentional properties of mental states; that is why, at the beginning

---

<sup>98</sup> Constitution/realisation by the physical doesn't entail supervenience on the physical as the following example indicates. Suppose that some physical objects have the property of being created by God. Such objects are physically constituted yet the property of being a physical object created by God does not supervene on the physical; rather it supervenes on factors outside the physical realm. Consequently, intentional phenomena could be physically realised without it following that intentional properties are supervenient on the physical. Therefore, to establish that intentional properties are scientifically respectable one needs to do more than determine that intentional phenomena are physically realised or constituted.

of this section, I described naturalists as having a worry that intentional mental phenomena do not belong to the natural world and that intentional properties do not have the right sort of reality. Who knows how the naturalisation story goes with respect to geological, biological, and perhaps even chemical properties. But few would doubt that there is a story to be told, no matter how complex the details. Thus only an eccentric would harbour any doubts about the scientific respectability of geology, biology or chemistry. Alas, the same cannot be said of intentional psychology, for philosophers who doubt that there is a naturalisation story to be told with respect to intentional properties are literally queuing up to tell us the bad news.<sup>99</sup> Thus we find Fodor making comments such as the following:

The worry about representation is above all that the semantic (and/or the intentional) will prove permanently recalcitrant to integration in the natural order; for example, that the semantic/intentional properties of things will fail to supervene upon their physical properties. What is required to relieve the worry is therefore, at a minimum, the framing of *naturalistic* conditions for representation. ('Semantics Wisconsin Style' p 32)

If the semantic and the intentional are real properties of things, it must be in virtue of their identity with (or maybe their supervenience on?) properties that are themselves *neither* intentional *nor* semantic. If aboutness is real, it must be really something else.

---

<sup>99</sup> Such philosophers make up a motley bunch. At one extreme there are those who, following Quine, believe that the semantic and intentional resist naturalisation as such properties constitute a closed circle and are happy to conclude: so much the worse for such properties and any descriptive and explanatory enterprise that trades in intentional talk. At another extreme there are those who, like McDowell and the later Putnam, think that the normative nature of meaning stands in the way of naturalisation, but that that is not to say that the intentional lacks reality or that all explanatory and descriptive enterprises that appeal to intentional properties thereby stand in a state of ill repute. Such philosophers object to what they see as the scientism of such philosophers as Fodor.

And, indeed, the deepest motivation for intentional irrealism derives from a certain ontological intuition; that there is no place for intentional categories in a physicalistic view of the world; that the intentional can't be *naturalised*. (*Psychosemantics* p 97)

Given these worries there is a very real need to make some progress in the task of constructing a naturalistic theory of content. Once we have such a theory the mystery as to how our mental states could have semantic and intentional properties, and any doubts as to the possibility of a scientifically respectable intentional psychology, will evaporate.<sup>100</sup>

If one is committed to the existence or the possibility of a scientific intentional psychology then one is thereby committed to the idea that the intentional properties that such a psychology appeals to in its theories and explanations are related to the physical in the very way that gives the naturalisation project its sense and importance. One cannot be a champion of scientific intentional psychology and deny that there is a naturalisation story to be told, a story that specifies just which aspects of the physical world fix or determine the intentional and semantic properties that that psychology ascribes to us and our mental states. Of course one is free to deny that there is, or could be, such a psychology. If one did this, one could consistently hold that there is neither a need for, nor a possibility of producing, a naturalistic theory of content. However, I don't find this position very attractive as it requires its advocate to hold either that contemporary cognitive psychology is not a legitimate science, or that its attributions of intentional properties to our internal states is not to be taken literally. Given that I have been arguing for a scientific cognitive psychology that, in the course of explaining our intentionally characterised cognitive capacities, makes literal and ineliminable attributions of intentional properties to some of our internal states, I am committed to there being a naturalisation story

---

<sup>100</sup> I think that there is something distinctive about semantic and intentional properties that gives some *prima facie* or intuitive plausibility to the idea that they are not related to the physical in the way that bona fide special science properties are. For example, intentional states can be about things that are external to them, things with which they have had no causal contact, and, indeed, things that do not exist.



to be told with respect to those properties; a story that needs to be told in the course of constructing a complete and successful vindication of that psychology.

### 6.3 Reductionism

What form should a naturalistic theory of content take? What kind of theory should the naturalist be seeking? One thought is that the theory should attempt to reduce the intentional to the physical in the manner in which the property of being water reduces to that of being H<sub>2</sub>O.<sup>101</sup> Much of the opposition to the project of naturalising content is based on the idea that to engage in that project is to thereby commit oneself to the idea that the intentional reduces to the physical.<sup>102</sup> Of course if intentional properties did reduce to physical properties it would be clear that (and how) intentional psychology could satisfy the requirements described in the previous section for being a respectable science. But is it really plausible that there is such a reductive relationship; that intentional properties have a non-intentional essence? Despite the well publicised examples of water, lightning, and heat, scientific investigation has uncovered few intertheoretic bridge laws and philosophers have proved even less successful in specifying convincing necessary and sufficient conditions for the application of even our most simple and familiar concepts. So we have every right to be sceptical about the prospects for naturalising intentional content by means of reduction.

The above quotes suggest that Fodor is seeking a reductionist theory. However, they are misleading, for, in keeping with his career-long opposition to the idea that the categories of the special sciences can (and must) reduce to those of physics, what he really seeks is a

---

<sup>101</sup> For a helpful account of the nature of the reduction relation see Kim (1996) ch. 9.

<sup>102</sup> See, for example, Stich (1992) and Tye (1992), both of whom represent the naturalisation project as being one of specifying reductionist necessary and sufficient conditions for the application of intentional concepts. Both are sceptical about the possibility of reducing the intentional to the physical, but deny that any disastrous consequences follow from this, given the fact that special science properties rarely reduce to the physical and need not so reduce to be scientifically respectable. In other words, they object to what they see as the reductionist assumptions underlying and driving the naturalisation project.

sufficient condition for meaning (as opposed to a necessary and sufficient condition). It is consistent with the existence of such a sufficient condition that intentional and semantic properties can have many and varied sources; that no such property has a non-semantic and non-intentional essence. The idea is that it is possible that my cow thoughts are about cows (rather than something else or nothing at all) for one reason, that the states of my visual system represent what they represent for some other reason, and that Fang's thoughts are about food for yet a third reason. In other words, in each of these cases some feature of the physical world is responsible for, fixes or determines, the intentional properties of the mental state in question; but that feature might differ from case to case.

But why, one might wonder, must there be any such non-semantic, non-intentional, sufficient conditions? The answer has to do with the doctrine of physicalism. Physicalism doesn't require that higher-level, special science properties reduce to those of physics, but it does require that where there are no relations of reduction or identity there are relations of realisation and supervenience. In other words, at the very least, higher-level special science properties are physically realised and have physical supervenience bases so that whenever such a property is tokened in our world the tokening of that property is the product of, or is fixed or determined by, some lower level physical feature of the world such that the higher level property in question couldn't but have been tokened given that physical feature.<sup>103</sup> Now the existence of such supervenience or realisation relations between higher-level special science properties and lower-level physical properties entails that associated with each special science property will be a battery of conditional statements each specifying a physical condition the satisfaction of which is sufficient for the tokening of the higher-level property in question. There may be infinitely many such physicalistic sufficient conditions associated with each higher-level property, each one specifying a distinctive way in which a tokening of that property can be generated. To deny of a higher-level non-physical property that there are any such

---

<sup>103</sup> In this context we need not worry about the specific nature of the supervenience relation (e.g. "strong", "global", or whatever) implied by physicalism, or whether the notion of necessity implied by "couldn't but have" is metaphysical or nomological.

physicalistic sufficient conditions associated with it is either to deny that it is scientifically respectable or to reject the doctrine of physicalism.

What these considerations imply about intentional properties is the following. If there is to be a scientifically respectable intentional psychology then the intentional properties that figure in the explanations and theories of that psychology must have physical supervenience bases. Thus, associated with each of these properties there must be physicalistic sufficient conditions such that the satisfaction of their antecedents guarantees the tokening of the property in question. To deny that there are such sufficient conditions is either to deny the possibility of a scientific intentional psychology or to reject physicalism. Anyone who is committed to physicalism (a doctrine that I take to be uncontestable and fundamental to the ideology of science) and to the reality or possibility of a scientific intentional psychology must accept that there are such physicalistic sufficient conditions for the tokening of intentional properties.

To attempt to uncover physicalistic sufficient conditions for meaning is to engage in the naturalisation project without thereby committing oneself to any implausible reductionist thesis. Seen in this light I think that it is clear that the naturalisation project is not a misguided project, and that it is an especially important one in which to make progress given the widespread scepticism concerning the prospects for a respectable scientific intentional psychology. Once we have specified some appropriate sufficient conditions we will have indicated that intentional properties supervene on physical properties and removed the mystery as to how intentional facts could be determined or fixed by physical facts.<sup>104</sup>

---

<sup>104</sup> Michael Tye (1992) argues that intentional psychology is a respectable science, and in the course of doing so constructs an account of the relationship between psychology (and higher level sciences in general) and physics that is similar to my account. He advances a version of what he calls "naturalism" according to which:

The general relationship which obtains between higher level and lower level physical properties is one of realization . . . The realization relation is, at least in part, one of determination: the lower level property synchronically fixes the higher level one so that the tokening of the former at any time *t* necessitates the token of the latter at *t* but not conversely. . . The parallel between types and

But what is the nature of the required physicalistic conditions? In *Psychosemantics* Fodor explicitly states that his theory is intended to specify a sufficient condition that is satisfied by our mental states. For example, he writes:

I want a *naturalized* theory of meaning; a theory that articulates, in nonsemantic and nonintentional terms, sufficient conditions for one bit of the world to *be about* (to express, represent, be true of) another bit. I don't care . . . whether this theory holds for *all* symbols or for all things that represent . . . I'm prepared, that is, that only mental states (hence, according to RTM, only mental representations) should turn out to have semantic properties *in the first instance*; hence that a naturalized semantics should apply, *strictu dictu*, to mental representations only.

But it had better apply to them. (pp 98-99)

However, by 'A Theory of Content II' it seems that Fodor has relaxed his ambitions somewhat so that he would be satisfied with any sufficient condition for meaning, even a condition that we, or

---

tokens on the above conception of naturalism should now be clear: higher level types may be realized by more than one type within the actual world; higher level tokens may be constituted by different lower level tokens but only in different possible worlds.

We are now in a position to summarize what naturalism with respect to mental states (token and type) comes to on the above account:

Mental states participate in causal interactions which fall under scientific laws, and are either ultimately constituted by or ultimately realized by microphysical phenomena. (p 436)

The main difference between Tye's position and mine is that he doesn't think that there are any necessary and sufficient, or purely sufficient physicalistic/naturalistic conditions for meaning. His point is that naturalism doesn't require that there be any such conditions. Hence, Fodor is engaged in a misguided project; intentional properties neither are, nor need be, naturalisable. However, if what I have said is correct, naturalism/physicalism does require that there be, at the very least, physicalistic sufficient conditions associated with any scientifically respectable intentional property. In denying the existence of such conditions Tye, leaves it a complete mystery as to how the intentional could be related to the physical in the way he describes.



our internal states, didn't satisfy. For example, he writes, 'It's enough if I could make good the claim that "X" would mean such and such if so and so were to be the case. It's not also incumbent on me to argue that since "X" does mean such and such , so and so is the case' (p 96).<sup>105</sup> I think that Fodor's initial ambition is the right one for the naturalist to have. Certainly to have a mere sufficient condition would be an achievement, and one relevant to the program of defending the scientific status of intentional psychology. The worry that there is no place for meaning in the natural world would be soothed were it shown that there would be meaning in the natural world if such and such (nomologically possible) nonsemantic and nonintentional conditions were satisfied. But the uncovering of a sufficient condition isn't going to indicate how *our* states could have semantic and intentional properties, or how there could be a scientific intentional psychology which was a psychology of *us* , if it isn't a condition that we (or our intentional states) satisfy.

It strikes me that - besides specifying a sufficient condition that we (or our internal states) satisfied - there are two further features that an adequate naturalisation story should have. The first feature is implied by what I wrote above. It is that the story should *explain* the semantic and intentional properties of our mental states in the sense of telling us in virtue of what our mental states have the semantic and intentional properties that they have. It should specify the supervenience bases of those semantic and intentional properties; tell us which features of the physical world fix or determine their tokening. To provide the sufficient condition Fodor envisages wouldn't automatically be to do this even if our states satisfied that

---

<sup>105</sup> And again:

Suppose we had naturalistically sufficient conditions for content. It wouldn't, of course, follow that any of our neural states, or any of our public symbols have the content that they do because they satisfy the conditions on offer. Indeed it wouldn't follow from the mere existence of sufficient conditions that anything in the universe has actually got any. . . On the other hand, if there are naturalistic sufficient conditions for content, and we don't know these conditions not to be satisfied, then we would at least be in a position to claim, for example, that "cat" could mean *cat* for all we know to the contrary. This would be a satisfactory situation for the philosophy of mind . . . to have finally arrived at. (p 131).



condition. This can be seen by considering the following case. I suppose that, in point of nomological fact, water is the only chemically pure substance that - under normal atmospheric conditions - has a boiling point of one-hundred degrees Celsius. Consequently, it's a sufficient condition for a sample of a chemically pure substance to be water that it boils at one-hundred degrees Celsius. But the water in my beaker isn't water in virtue of its boiling point; rather it is water in virtue of its molecular structure.

One could imagine Fodor uncovering a property analogous to that of having a boiling point of one-hundred degrees Celsius, a property that our mental states had in virtue of their having the properties that fix or determine their meaning in the manner in which one's molecular structure determines one's boiling point. A sufficient condition that appealed to this property wouldn't explain or tell us in virtue of what our mental states have the semantic and intentional properties that they have.<sup>106</sup>

A second desirable feature of any naturalisation story is that it should be able to deal with all of the representational states and semantic and intentional properties that feature in psychological theory and explanation. Fodor concentrates his attentions on such familiar propositional attitudes as beliefs and desires aiming to account for the contents of such personal level states. This is because of his view of scientific psychology as being an extension of folk psychology. Now if beliefs and desires have some role to play in scientific psychology, then to construct a theory that naturalised their contents would count as a significant achievement. But if that story didn't apply to the sub-personal representational states that cognitive psychologists routinely attribute to us and our cognitive modules

---

<sup>106</sup> This point indicates how we should respond to an objection to Fodor's project made by Stich (1992). Stich argues that Fodor needs more than sufficient conditions to naturalise semantic and intentional properties, since 'providing conditions that are merely sufficient is just too easy' (p 363). He cites as an example of an easily generated sufficient condition the following: 'If  $x$  is Fodor's most recent utterance of "Maria Callas" (or: if  $x$  is the concept that underlies that utterance) then  $x$  represents Maria Callas' (p 363). The obvious reply is that this condition does not tell us in virtue of what  $x$  represents Maria Callas or which feature of the physical world fixes or determines the meaning of  $x$ . A sufficient condition that does that is certainly not too easy to generate.

(and if the advocate of that story had no supplementary theory to deal with such states), then the scientific psychology that we actually have will hardly have been vindicated. For it will not have been shown that, and how, the properties that it appeals to have a firm place in the natural order.

This second requirement is not an idle demand, given the (vaguely Wittgensteinian) thought that the nature of meaning is such that there couldn't be meaning at the sub-personal level. Here's how the thought goes. An item is meaningful only if its use is governed by a system of rules of a linguistic community, rules which determine not just how the item is in fact used but also how it should be used. To be used in a meaningful way - for example to represent or communicate a fact about the external world - the item must be employed to do so by an agent who belongs to the linguistic community in question, grasps its rules, and succeeds in following them.<sup>107</sup> Now what results do we get if we apply these considerations to the sub-personal? Neither my brain nor any subsystem of it, is an intelligent agent that belongs to a linguistic community, grasps its rules, and employs the symbols of that community's language in a manner that constitutes rule following. Consequently, whatever my brain or neural subsystems are doing it is not engaging in semantic or linguistic activity. Thus a cognitive psychology that seeks to uncover and describe our internal workings cannot legitimately talk of representation or meaning, nor attribute semantic and intentional properties to our sub-personal states in any literal or full blown sense. It can only ascribe semantic and intentional properties in a non-literal, instrumental, or metaphorical respect, or, as McDowell (1994) might say, it can only attribute *as if* content. Therefore there cannot be a legitimate intentional psychology of our sub-personal workings.<sup>108</sup>

---

<sup>107</sup> Where to follow a rule is to do more than merely act in accord with it. The nature of the difference between following a rule and merely acting in accord with it need not concern us here. Suffice it to say that Wittgensteinians feel that in order to follow a rule, at a minimum, one has to grasp that rule, something that requires intelligence and understanding. One can, on the other hand, act in accord with a rule whilst being completely bereft of intelligence and thus without grasping it.

<sup>108</sup> A related argument that is generated by these kinds of considerations is that an intentional cognitive psychology would be explanatorily useless in virtue of being

Given such objections the naturalist (or at least the naturalist sympathetic to an intentional cognitive psychology of the form that I have attempted to describe in earlier chapters) needs to tell us how there could be meaning at the sub-personal level; how, for example, states of the human visual system could have semantic and intentional properties in any literal, full blown respect. The upshot of all this is that if Fodor's sufficient condition only applies to personal level intentional states then he will not have succeeded in naturalising all of the intentional and semantic properties that are appealed to in contemporary cognitive psychological theory and explanation. He will have left it a mystery how, for example, Marr's 3-D model representation could represent the shape of an object impinging on a human subject's visual system. Indeed, I do suspect that the naturalisation story with respect to such states will be different from that with respect to beliefs, desires, and the like; that there will be no single sufficient condition that is satisfied by all our intentional states.

---

either circular or leading to infinite regress by attributing to our brains the very capacities for which an explanation is sought. This point is related to the homunculus argument described by Dennett (1978), and can be outlined by reference to an example Wittgenstein presents in *The Blue Book*. Wittgenstein asks how we obey the order "fetch a red flower". An answer to this question that Wittgenstein considers is that we compare the flower before us with a mental image of red, picking the flower if, and only if, it matches the image. One problem with this, he argues, is that if it requires to be explained how we are able to perform the task in question then it also requires to be explained how we are able to select the correct mental image. The phenomenon that the putative explanation appeals to is just like that which it is invoked to explain, so making the explanation circular. And if the advocate of the putative explanation attempts to explain the capacity to select the correct mental image by appeal to a process of comparing the mental image with a prior mental image of red then an infinite regress looms. Thus the invocation of a mental image does no real explanatory work and any appearance to the contrary is an illusion that is the product of the idea that the mental image resides within that mysterious occult medium called the mind.

## 6.4 Fodor's method

So we need a naturalistic theory of content, a theory that specifies - in nonintentional and nonsemantic terms - a sufficient condition (or a battery of sufficient conditions) for the tokening of those intentional and semantic properties attributed to our internal states in the course of psychological theory construction and explanation: a sufficient condition (/battery of sufficient conditions) that, moreover, is (/are) satisfied by our psychological states and which explains why they have the semantic and intentional properties that they have. But how are we to construct such a theory? Is a priori reflection the right method to adopt or is a more empirical approach needed? Indeed, is the task philosophical at all? Fodor's approach is decidedly a priori and is reminiscent of the traditional technique of conceptual analysis. He engages in no examination of psychological or scientific research, instead proposing a sufficient condition from his armchair, and modifying that condition in the light of any counterexamples he can come up with or is presented with.

Stich (1992) claims that many philosophers who engage in the naturalisation project employ the method of conceptual analysis. What they attempt to do is construct definitions of, or uncover necessary and sufficient conditions for the application of, our semantic and intentional concepts from their armchair by proposing definitions of the target concepts, searching for counterexamples to them, and modifying the definitions in the light of any counterexamples so discovered.

Stich thinks that this employment of the method of conceptual analysis is thoroughly misguided as it is based upon the incorrect assumption that all of those who possess a mastery of intentional and semantic concepts have, underlying that mastery, tacit knowledge of necessary and sufficient conditions for the correct application of those concepts. Stich argues that cognitive psychological research, in particular the work of Eleanor Rosch<sup>109</sup>, suggests that this assumption is likely to be mistaken. Rosch's work suggests that there is no tacit knowledge of necessary and sufficient conditions for the application of our concepts underlying our classificatory practices. I regularly, and correctly, assign objects that I

---

<sup>109</sup> See, for example, Rosch 1973, 1975 and 1978.



perceive to the category "bird". But underlying this capacity to classify birds correctly is not a representation or mental structure that explicitly represents necessary and sufficient conditions for being a bird. Rather the representation implicated is a prototype, that is, an idealised description of prototypical members of the category "bird". Whenever I seek to determine whether an object before me is a bird, a similarity metric is employed so that I will judge that the object is a bird if it resembles the prototypical bird (as represented in me) to a sufficient extent. The prototypical features of birds explicitly represented in me might be those of having feathers, of being able to fly, and such like. What is important to note is that these features constitute neither necessary nor sufficient conditions for being a bird.

Stich induces from such psychological research the conclusion that the representations or mental structures with respect to such intentional categories as believing that *p* and desiring that *q* do not explicitly represent necessary and sufficient conditions for belonging to them. He writes: 'Perhaps the safest bet is that, whatever the mental mechanism underlying intentional categorization may be, it will not utilize "classical" concepts - the sort that can be defined with a set of necessary and sufficient conditions' (p 353). Thus, he concludes, it is thoroughly misguided to engage in armchair conceptual analysis in order to lay bare the essence of the intentional and semantic properties that we ascribe to the states of ourselves and our fellows.<sup>110</sup>

What bearings does the above argument have on Fodor's project and the viability of the method he employs? We have seen that Fodor is not in the business of defining or laying bare the essence of the intentional and semantic concepts we employ (or the properties they express). But still, it might be thought, Fodor assumes that there are sufficient conditions for the application of our intentional and semantic concepts, an assumption that Stich's argument suggests to be mistaken.

It strikes me that Stich is mistaken in claiming that the psychological evidence suggests that our concepts are not "classical", that they cannot be defined by means of necessary and sufficient conditions. Stich's mistake is to fail to recognise that there is a

---

<sup>110</sup> Similar arguments for this conclusion, that again appeal to the work of Rosch, are presented by Tye (1992) and Horgan (1994).



difference between what philosophers (and ordinary folks, come to that) are talking about when they use the term "concept" and what psychologists mean by the very same term. In the technical psychological sense a concept is a mental representation. In the philosophical sense a concept is not a mental representation but rather an abstract object that we grasp and apply in the course of our mental life.<sup>111</sup> This is, of course, not to deny that there are mental representations within us that underlie our grasp of concepts, and that are implicated in our employment of them.

Now those who engage in conceptual analysis are invariably interested in concepts as opposed to mental representations. What Rosch is concerned with, on the other hand, is the nature of the representations that underlie categorisation. If she has succeeded in establishing that these representations do not explicitly represent necessary and sufficient conditions she hasn't thereby shown that our concepts can't be defined by means of necessary and sufficient conditions or that there are no such conditions for their application. Maybe what hooks me, so to speak, onto the concept *bird* is a complex representation that carries information about birds but does not carry a definition of what it is to be a bird. But that doesn't entail that an elucidation of the concept that I grasp would be a specification of that information. The representation that hooks other people onto that same concept may well be quite different; indeed it hardly seems plausible to suggest that the prototypes underlying our classificatory activities are invariant across the population of individuals who are capable of making the same judgements as to what is before them. This further indicates that concepts cannot be identified with mental representations. Indeed it is hard to see how Rosch's claims could be coherently expressed without assuming that the concepts that we apply in categorising objects and the mental representations implicated in such activities are not one and the same. Thus claims about the nature of the concepts that we grasp and apply cannot be automatically read off from facts about the nature of the mental representations underlying our grasp and employment of those concepts. But that is precisely what Stich does.

---

<sup>111</sup> From this point on I will use the term "concept" in the philosopher's sense and the term "mental representation" to refer to what psychologists mean by "concept".

Of course none of the above implies that our concepts can be successfully analysed, and it might be argued that the track record of conceptual analysis hardly gives grounds for optimism. However, I sympathise with Crispin Wright (1989) when he says: 'it seems to me that it is an important methodological precept that we do not despair of giving answers to constitutive questions too soon; if the accomplishments of analysis in philosophy often seem meagre, that may be because it is difficult, not impossible' (p 246). So I don't think that to search for necessary or sufficient conditions for the application of our concepts is necessarily to engage in a hopeless and misguided project.

But is a priori reflection the appropriate method for uncovering any such necessary and sufficient conditions? Doesn't the work of psychologists such as Rosch indicate that even if our concepts are definable by means of necessary and sufficient conditions we do not have any knowledge of such conditions, thus implying that empirical investigation is the only sensible way of proceeding? Perhaps not, for to engage in analysis is not to attempt to make explicit the tacit knowledge that we rely upon in making judgements as to the applicability of our concepts (or to make explicit the information encoded in the representations that underlie such judgements). So those who engage in analysis do not thereby operate on the assumption that all of those who grasp a given concept tacitly know any necessary and sufficient conditions for its application. Thus, once again, Rosch's discoveries do not tell against their activities.

It strikes me as not absurd to think that if a concept is such that a constitutive account of it could be given, then that account could be constructed on the basis of a priori reflection. Suppose that I have a grasp and mastery of a concept C. I often apply this concept in the course of my daily life, and typically such applications are correct. Aren't I then in a good position to use my faculty of reason to work out what binds together those cases where C can be applied and what marks them apart from those cases where an application of C would be incorrect? Couldn't I, by a process of tentative hypothesis formation, a search for counterexamples, and subsequent hypothesis modification, tease out a constitutive account of C, or necessary and sufficient conditions for its application, utilising my tacit knowledge

in the process? To do this wouldn't be to search for what one tacitly knew at the beginning of the process, but rather to arrive at new knowledge, knowledge that most of those with a mastery of C do not possess (even tacitly). In short, to suppose that one could generate a constitutive account of a concept by means of a process of a priori reflection that involved utilising what one tacitly knew at the beginning of the process is not to commit oneself to the idea that one had knowledge of (if only tacit knowledge) that constitutive account all along. So, at least with respect to some of our concepts, the search for a constitutive account of them (or necessary and sufficient conditions for their application) by means of conceptual analysis and a priori reflection may not be as misguided as Stich would have us believe. But what are the concepts of which there might be a constitutive account? Why, those that philosophers have traditionally been interested in, such as those of truth, justice, knowledge, beauty and, perhaps, such semantic concepts as meaning, reference, and so on.

However, none of this should be taken as constituting an endorsement of Fodor's a priori approach for two reasons. Firstly, to accept the in principle possibility of constructing a constitutive account of some of our intentional and semantic concepts by means of a priori reflection is not thereby to say that that account would be naturalistically acceptable. Maybe the only constitutive account to be had in this area is semantic and intentional through and through, an account that elucidates our intentional and semantic concepts in irreducibly semantic and intentional terms by way of a specification of the intricate relationships between these concepts. Thus, my recent reflections do not involve an endorsement of the reductionism that I earlier cast doubt on. Secondly, the naturalist's project, as we have seen, is not that of generating a constitutive account of our semantic and intentional concepts at all. Rather, it is that of discovering naturalistically acceptable sufficient conditions for the tokening of the semantic and intentional properties attributed to our states in the course of psychological explanation and theory construction; conditions that indicate how those properties are realised in us, what their supervenience bases are, or what features of the physical world fix or determine them when they are tokened. Given the nature of the naturalist's goal, whatever one thinks about

the powers of conceptual analysis and a priori reflection, one may well suspect that they are singularly ill-suited as tools for an effective completion of the naturalisation project. After all, with respect to many of the properties appealed to by scientists, it would be very odd to suppose that anything other than a good deal of empirical investigation could reveal how they are realised in our world, what their supervenience bases are, or what features of the physical world fix or determine them when they are tokened.<sup>112</sup>

However, on the other hand, all is not lost for a priori reflection for there is a significant difference between many intentional and semantic properties and some of the more esoteric scientific properties. We ascribe intentional and semantic properties to the states of ourselves and our fellows all the time and we are very effective in making correct ascriptions. Thus we are, perhaps, in a good position to reflect upon what underlying facts generally hold when a particular intentional property is tokened, and so generate tentative hypotheses as to what the required sufficient condition comes to. In other words, perhaps we can tease out a satisfactory naturalistic theory of content by engaging in a process analogous to that described above for constructing a constitutive account of those concepts that have traditionally been of special interest to philosophers.

There is a further reason for thinking that a priori reflection has a significant role to play in the achievement of the naturalist's goal. If we are going to get clear on what determines the semantic and intentional properties of our psychological states it is important that we have a good understanding of the nature of those properties. Quite generally, we need to know what it is that we are attempting to naturalise if we are going to have much chance of naturalising it. If we do not have this kind of understanding there will be a real danger that we will rest satisfied with sufficient conditions that work with respect to our intentional and semantic properties as we take them to be, but not as they are in reality. I take it that a priori reflection will

---

<sup>112</sup> For example, if you want to know the answer to such questions as to how genetic properties are realised in us, what their supervenience bases are, and what features of the physical world fix or determine them when they are tokened, my advice to you is that you go and consult a scientist. As enthusiastic as I am about my subject, I don't think a philosopher will be much help.



play an important role in achieving this understanding by enabling us to uncover the relations between our various semantic and intentional concepts, by, for example, answering such questions as whether there is more to meaning than reference, whether normativity is an essential component of meaning, and so on. This point suggests the following criticism of Fodor's approach. Fodor dives headfirst into the naturalisation project, engaging in little by way of investigation into the nature of the properties that he attempts to naturalise. Consequently, he has to deal with objections to his theory according to which it is unable to deal with, or ignores, some important aspect of meaning (such as sense or normativity) in what often looks like an ad hoc manner. Perhaps a more cautious and circumspect approach would be wiser and more fruitful in the long run.

Yet although a priori reflection no doubt has an important role to play in giving us the required understanding of the semantic and intentional properties that we seek to naturalise, a priori reflection alone will not be enough. We are forever making judgements as to the contents of the propositional attitudes of ourselves and our fellows; we are very familiar with the semantic and intentional properties that such mental states as beliefs, desires, and the like, have. But there is a whole range of intentional states of which most ordinary folk are ignorant, namely the sub-personal states that cognitive psychologists are so concerned with. There is no guarantee that the kinds of semantic and intentional properties that these states have will match up with the familiar properties of beliefs and desires. Indeed, given the obvious differences between these two broad categories of mental states, there is every reason to suspect that they diverge quite considerably in their semantic and intentional properties. Given that the naturalist has to deal with subpersonal intentional states just as much as she has to deal with beliefs and desires, a full understanding of the nature of the semantic and intentional properties that require to be naturalised cannot be acquired without consulting the psychologist. To proceed with blithe ignorance of, and indifference to, psychological research is to court with disaster.

There are further reasons why a consideration of psychology is important. Perhaps we are all in a position to detect from our



armchair some underlying conditions that are invariably satisfied whenever an individual believes that, for example, there is a horse before her. By considering such conditions and reflecting on the question of whether they could play the role of determining meaning we might be able to work out the relevant naturalistic story with respect to such states. But most of us are very ignorant of the facts surrounding the tokening of subpersonal intentional states. These states may well get their contents fixed in a way quite different from that in which beliefs and desires get their contents fixed. So if we are going to produce a naturalistic theory that applies to such states, we are going to have to turn to psychology for help.<sup>113</sup>

Moreover, even a naturalist who seeks an account that applies only to beliefs and desires would be well advised not to ignore psychological research. A putative naturalistic sufficient condition might look convincing if we restrict our attention to beliefs and desires in virtue of its having an air of intuitive plausibility and our not having discovered any counterexamples. However, there might be counterexamples aplenty within the subpersonal realm; cases of states that satisfy the condition in question yet clearly don't have the relevant semantic or intentional property; counterexamples that we would have soon become aware of if only we had turned our attention away from a narrow range of familiar beliefs and desires. This point is not merely speculative for in the next chapter I will criticise certain aspects of Fodor's theory by appeal to the kinds of states that cognitive psychologists have postulated.

What all this suggests is that Fodor's method of proceeding is not entirely appropriate given the nature of the task at hand. A priori reflection has an important role to play in the uncovering of the range of sufficient conditions that will collectively naturalise the target intentional and semantic properties. But to engage in a priori reflection whilst remaining ignorant of, and indifferent to, developments in scientific psychology (and perhaps other sciences as well) is to adopt a method that does not inspire my confidence. Surely what is required is a more inter-disciplinary approach.

---

<sup>113</sup> Indeed we may well have to turn to sciences other than psychology to get any indication as to what the relevant content determining facts might be; neuroscience, for example.

## 6.5 Normativity

In certain quarters it is thought that the central obstacle that stands in the way of a successful completion of the naturalisation project - and, indeed entails that that project is ill-conceived - is the normative nature of meaning. Given that I have been attempting to vindicate the naturalisation project I cannot ignore this important and influential line of thought. I cannot hope to do anything more than scratch the surface of the complex and difficult issues in this area. However, I think I have a few points to make that suggest that normativity doesn't constitute as big a problem for the naturalist as some would have us believe.

What is involved in the claim that normativity is central to meaning or that meaning is a normative notion? The basic idea is that normativity is an essential component of meaning in the respect that if a symbol has a meaning then there will be infinitely many facts concerning how it ought to be used in various possible circumstances over and above any facts concerning how it would be used in those circumstances. How a symbol ought to be used is one thing, and how it is in fact used (or would be used) is quite another; there is no guarantee that the two will coincide. And how a symbol ought to be used is partly constitutive of its meaning so that to attribute to a symbol a particular meaning is to imply a whole battery of normative facts.

This alleged normative element of meaning poses a challenge to the naturalist for it entails that, for her theory to be adequate, the nonintentional and nonsemantic properties or facts specified by it must be capable of determining or fixing the normative facts concerning how the symbol or state in question ought to be used. Now the worry is that nonsemantic and nonintentional properties or facts are just not capable of fixing or determining such normative properties and facts. Consequently, there is no true naturalistic theory of content and to attempt to construct such a theory is to engage in a misguided project. In short, the objection is that the normative nature of meaning takes it out of the natural, nonsemantic realm.<sup>114</sup>

---

<sup>114</sup> This line of thought is closely associated with Kripke's Wittgenstein (Kripke, 1982) and with the work of John McDowell (see, for example, McDowell, 1986 and 1994a and McCulloch, 1995). Of course, a fundamental difference between Kripke's

I do not find this objection convincing; the naturalist can deal with normativity, and that she can has a lot to do with the fact that there are meaningful items that do not have a constitutive normative element. In what follows I will try to substantiate this claim.

A first point worth making is that the naturalist is not required to reduce normativity, or normative properties and facts, to physical properties and facts. It is quite consistent with naturalism that the normative is irreducibly normative/semantic. All that requires to be shown is that (and how) such normative facts as there are are ultimately determined or fixed by the physical.

It might be thought that there is no problem of normativity over and above the problem of misrepresentation; that to say that to use a symbol in a particular way is to use it in a way it ought not be used is to say no more than that such a use would constitute a case of misrepresentation. Now those philosophers who have attempted to construct a naturalistic theory of content have focused much of their energies on dealing with the problem of misrepresentation; with accounting for, in nonsemantic and nonintentional terms, the difference between tokenings of symbols that correctly represent their cause and tokenings of symbols that misrepresent their cause. It is far from obvious that such attempts to deal with the problem of misrepresentation fail, or that they must inevitably fail. Hence, if the problem posed by the normativity of meaning is nothing other than the problem posed by the phenomenon of misrepresentation then there are no compelling grounds for a blanket scepticism regarding the prospects for a viable naturalistic theory of content. However, it would appear that the problems of normativity and of misrepresentation are not one and the same. To say that a system has produced a symbol that misrepresents that symbol's cause is not to imply that the system has done anything it ought not to have done. Once a naturalist has successfully dealt with misrepresentation she

---

Wittgenstein, on the one hand, and McDowell, on the other, is that the former believes that the inability of the nonsemantic and nonintentional to fix the normative entails that there is no such thing as meaning something by a term, whereas the latter does not. For McDowell the presence of meaning in our world does not require that there be a deeper, more basic level of nonsemantic facts that are capable of generating, or being responsible for, that presence.

still owes us an explanation of what makes misrepresentation something that a system ought not do.

It strikes me as false to say that normativity is an essential component of meaning, for there are cases of meaningful items with respect to the use of which there are no normative facts. Consider the visual module as described by Marr. It produces symbols that certainly have semantic and intentional properties. For example, the primal sketch represents the nature and location of sudden intensity changes in the retinal image, the 3-D model representation represents the shape of objects impinging on the subject's visual apparatus, and so on. Yet governing the generation of these symbols are no norms or rules determining what the visual module ought (and ought not) do in various possible circumstances. It would be very odd indeed to say that my visual system had done something it ought not do when it produced a 3-D model representation that misrepresented a square shaped object as being rectangular; just as odd as to describe a plant that developed buds during a pre-spring warm spell as having violated some norm governing its budding behaviour. I think that this point generalises to all symbols tokened at the sub-personal level, that is to all representations that express the contents of sub-personal intentional states. Moreover, if Fodor's language of thought hypothesis is correct (so that symbols of a neurophysiologically realised language expressed the contents of our beliefs, desires, and the like) there would be no norms associated with the use of the symbols of that language, even though they expressed the contents of personal level, as opposed to subpersonal, states.<sup>115</sup>

---

<sup>115</sup> McGinn (1984) might appear to be making the same point when he writes:

The issue of normativeness . . . has no clear content in application to the language of thought: what does it mean to ask whether my current employment of a word in my language of thought (i.e. the exercise of a particular concept) is *correct* in the light of my earlier employment of that word? . . . There is just no analogue here for the idea of linguistic incorrectness (as opposed to the *falsity* of a thought): linguistic incorrectness . . . is using the same word with a different meaning from that originally intended (and doing so in ignorance of the change), but we cannot in this way make sense of employing a concept with a different



None of this is to deny that there are meaningful items such that there are normative facts closely associated with their meanings. For there are certainly norms governing the use of public language symbols. When I apply the English word "horse" to a cow, use a word contrary to its meaning in my linguistic community, or am inconsistent in what I mean by a word, I do something that I ought not do. And as thinking sometimes involves the use of public language symbols, as when I say to myself "that dog looks ferocious", some of our thinking is norm governed. So there would appear to be a crucial difference between public language symbols on the one hand, and symbols of the language of thought and those manipulated by our subpersonal processors, on the other. What could be responsible for this difference?

Here are some salient facts about public language symbols. Public language symbols have a meaning in a linguistic community, a meaning that is dependent upon the intentional states of the members of that community (their collective beliefs, desires, intentions, and so on), but that is independent of the intentional states of any particular member. There are rules governing the use of these symbols that are similarly dependent on those intentional states, rules that the members intend to grasp and follow and that they demand that their fellows follow. An individual's use of the symbols she uses is mediated by an understanding of their meaning. Associated with those symbols and their use is a whole battery of intentional states, such as beliefs as to the meaning of the symbols, desires and intentions to use them in accord with their meaning in the wider community, intentions to use them consistently, and so on. Consequently, whenever an individual uses a symbol to mean something at odds with its meaning in the wider community, violates a rule governing the use of a symbol, or is inconsistent in her use of a symbol, she thereby violates or fails to satisfy her intentions with respect to that symbol. Moreover, she runs the risk of

---

content from that originally intended - it would just be a *different concept*. (p. 147)

However, I don't want to endorse McGinn's reasoning for his conclusion. As will become clear, I think that it is for quite different reasons that the use of the symbols of the language of thought is not norm-governed.



provoking the censure of the wider linguistic community that makes demands of her parallel to those that she makes of herself.

With respect to the symbols of the language of thought, and those manipulated at the subpersonal level, matters are somewhat different. They are not symbols of a shared public language governed by rules determined by the intentional states of their users. Neither are their meanings understood or grasped by their users. And their users do not have any intentional states with respect to their use.

My thought is that these differences are responsible for, and explain, the fact that there are normative facts associated with the use of public language symbols, but none with respect to symbols of the language of thought and symbols manipulated at the subpersonal level. This suggests a way of naturalising normativity. Such normative facts as there are can be accounted for by appeal to the intentional states of individuals; for it is my intentional states, along with those of my fellows, that determines the normative facts with respect to my use of the words and expressions of English. So long as these underlying intentional states do not themselves involve the use of symbols whose use is similarly surrounded by intentional states (or so long as any such surrounding intentional states can ultimately be dealt with by appeal to intentional states underlying them that satisfy this condition) then circularity or infinite regress is avoided. These underlying intentional states can then be naturalised by appeal to properties and facts that are not directly required to account for any normative properties or facts.

It might be objected that this appeal to linguistic rules along with our intentions to follow them and our practice of criticising our fellows when they violate them, can at best only account for normative facts connected with the use of public language symbols. But, so the objection continues, there are norms of theoretical and practical reasoning quite apart from linguistic rules. For example, there are rules or standards of inductive reasoning that I violate when I make a sweeping claim about the Finnish national character on the basis of meeting one or two Finns. In jumping to my conclusion from the premises from which I began, I have done something that I ought not have done and that this is the case has nothing to do with my having violated a rule of English that I

intended to follow. How is the naturalist to account for such normative facts?

In fact, I think that my treatment of language suggests a way of dealing with the present problem. There are many rules of inference and it is an open empirical question which of these our thinking employs or conforms to. As a matter of empirical fact, we place a very high value on truth and knowledge. We desire and seek to know, to hold beliefs that are true, to purge our belief system of false beliefs, and to make inferential leaps that have a high probability of taking us from true beliefs to true beliefs. Moreover, we criticise the ignorant and the indifferent, those who hold false beliefs, and those who make inferential leaps that have little chance of getting them from true beliefs to true beliefs (not to mention those who attempt to deceive their fellows). This is all to the good for having knowledge, holding true beliefs and minimising false beliefs generally aids our survival and helps in the satisfaction of our goals, ends and purposes. Consequently, when I reason badly, (as in the above example of jumping to a rash conclusion about Finns) I have acted in a way that runs counter to much of what I hold dear, in a way that violates some of my most fundamental intentions and desires. Moreover, I have done precisely the kind of thing that evokes the criticism and condemnation of my fellows as reasoning is judged partly by reference to the extent to which it conforms with well established and entrenched modes and rules of inference. My point is that such facts about our desires, intentions, and what we hold dear, along with facts about our critical and evaluative practices, are collectively responsible for the normative facts connected with theoretical and practical reasoning. In other words, just as in the natural language case, we can account for such normative facts as there are in connection with reasoning by appealing to intentional states and critical and evaluative practices bound up with those intentional states.

This line of thought could do with much in the way of extension and elaboration. But I think that I have done enough to suggest that - in virtue of the fact that normativity isn't an essential component of meaning - the problem of normativity doesn't present an insurmountable obstacle to the successful execution of the naturalisation project.

## 6.6 Conclusion

In this chapter I have attempted to develop an account of the nature of the naturalisation project that reveals it *not* to be a misguided project based upon false or confused assumptions. In particular, it is not an inherently or inevitably reductionist project. If there is to be a scientifically respectable intentional psychology then the phenomena that falls within its domain of enquiry must be physically composed, constituted, or realised, and the intentional and semantic properties that figure in its theories and explanations must supervene upon physical properties, or have their tokenings fixed or determined by physical facts. To engage in the naturalisation project, to attempt to construct a naturalistic theory of content, is to endeavour to indicate that and how the relationship between the psychological and the physical satisfies this condition, and thus vindicate scientific intentional psychology. A satisfactory naturalistic theory of content will consist of a battery of nonintentional and nonsemantic sufficient conditions that collectively indicates that and how the semantic and intentional properties of all of those intentional states of concern to scientific psychology are physically fixed or determined. The method most appropriate for constructing such a theory is interdisciplinary in nature, having both empirical and a priori components.

In the next chapter my attention will focus on Fodor's naturalistic theory of content.

## Chapter 7

### Fodor's Theory of Content

#### 7.1 Introduction

In Chapter 6 I argued that we need a naturalistic theory of content, a theory that specifies the nonintentional and nonsemantic determinants of the semantic and intentional properties attributed to our mental states in the course of psychological explanation and theory construction. So the question arises of whether Fodor's theory adequately performs this task. In this chapter I will address this question arguing in favour of a negative answer. I will also, in the light of my earlier reflections, attempt to draw some morals as to how we should proceed in the light of the failure of Fodor's theory.

Fodor's approach to the naturalisation problem is to a large extent dictated by his commitment to the Representational Theory of Mind. RTM entails that the task of specifying the determinants of the semantic and intentional properties of our intentional states reduces to that of specifying the determinants of the semantic and intentional properties of sentences of the language of thought or Mentalese. And, given that Mentalese has a combinatorial semantics, that reduces to the task of specifying the determinants of the meanings of the words of that language. Fodor's theory takes the form of a sufficient condition for a word of Mentalese to have the meaning that it has. In the next section I will turn to the task of describing that sufficient condition.

#### 7.2 Fodor's sufficient condition

Fodor recognises that different types of symbols have their meanings determined in different ways and thus that one single account will not apply to all the symbols of Mentalese. He proposes a functional/causal role theory for the logical symbols of Mentalese. And, echoing Kripke (1972/1980) and Putnam (1975), he proposes a causal chain story for proper names. Neither of these stories are told

in any detail,<sup>116</sup> reflecting his view that the hardest, and most important, part of the naturalisation project is that of dealing with those words which are neither logical symbols nor proper names; words, that is, that express properties, words like the English words "red", "horse", "proton", "virtue", and the like. In short his sufficient condition is supposed to apply to the simple, nonlogical symbols of Mentalese.

The intuition driving Fodor is that the meaning of a simple, nonlogical symbol is determined by its causal relations to phenomena external to it. Thus, for example, the Mentalese symbol HORSE means *horse* - or, in other words, expresses the property of being a horse - because horses, and only horses, reliably cause tokenings of HORSE.<sup>117</sup> Or because it's a law that horses cause tokenings of HORSE. The theory that Fodor offers is a modification of such a causal covariation theory, thus placing him in a long tradition of tying meaning to etiology. (Other members of this tradition include Dretske (1981) and Skinner (1957)).

The causal covariation theory is atomistic as, according to it, a symbol's meaning is determined not by its relations to other symbols but wholly by its causal relations to whatever property it expresses. Thus the theory countenances the possibility of a subject's having *horse* thoughts despite being unable to have any other kind of thought. This feature of the theory explains a lot of its appeal to Fodor for, as we saw in Chapter 5, he thinks that the truth of holism would imply the impossibility of intentional psychology. Given that the motivation for naturalising content is to vindicate intentional psychology, holist theories of content are not an option for him: hence his opposition to the functionalist, conceptual role theories of Block (1986), Harman (1982) and Loar (1981).

Quite apart from its atomism, there are several reasons for taking the causal covariation theory seriously; its widespread and perennial appeal is not without explanation. Firstly, by invoking causal relations and causal laws the theory appeals to something that is

---

<sup>116</sup> In fact he more or less tells them in passing.

<sup>117</sup> I shall refer to representations by means of capitalised words and expressions, meanings by means of italicised words and expressions, and natural language symbols by means of words and expressions flanked by inverted commas.



naturalistically kosher. And, one might wonder, what else has the naturalist got to play with? Secondly, a cursory look at the predicates of natural language would suggest that there is a close relationship between the meanings of such symbols and the causal relations they bear to the properties they express. For example, most competent, adult speakers of English are disposed to respond to horses by uttering "horse" and we are reluctant to attribute a grasp of the meaning of that word to children or foreigners who indicate that they are not so disposed, or who are disposed to utter "horse" in response to non-horses. If there is a close relationship between the meaning of the predicates of natural language and the causal relations that they bear to the properties that they express then there is every reason to expect that the same will be true of their Mentalese analogues. For, whenever we respond to an external object by uttering "horse" underlying that linguistic act will be a thought involving a tokening of HORSE. Thirdly, if a creature is to survive and prosper it must correctly represent the world external to it. Being able to think and reason will be of limited survival advantage if one cannot perceive the world around one. This implies that the representations produced by the perceptual systems of successful creatures will generally be caused by, and only by, instances of the property that they represent. Given the centrality of perception to mental life, it is tempting to think that this fact about representations could play a meaning-determining role.

However, the causal covariation theory faces a major problem, namely that of finding a place for error or misrepresentation. It is one thing to say that minded creatures have, and must have if they are to survive, very effective perceptual systems and quite another to say that they have perceptual systems that never get it wrong, that never misrepresent the nature of their distal stimulus. Indeed we all misrepresent the world from time to time as in the case where I see a cow on a dark night and mistakenly conclude that it is a horse. In such a case a cow causes me to token HORSE. The causal covariation theory seems to rule out such a possibility, implying that as cows on a dark night, as well as horses, cause tokenings of HORSE that symbol must mean something like *horse-or-cow-on-a-dark-night*. So, on the face of it, it looks like the theory doesn't apply to us. In the light of such a problem the advocate of the causal covariation theory needs to

modify the theory in such a way as to avoid being forced to make a mistaken attribution of a disjunctive content in such cases. In Fodor's terminology, what is needed is a solution to the disjunction problem.

To make matters worse, as Fodor emphasises in 'TOCII', there is another familiar kind of case where the cause of a Mentalese symbol does not fall within its extension, namely representation in thought. For example, thoughts about hay sometimes cause thoughts about horses; and thus sometimes cause tokenings of HORSE. When this happens we don't have error or misrepresentation; when a thought about hay causes me to token HORSE, I haven't mistaken the hay thought for, or misrepresented it as, a horse. Such cases seem to force the advocate of the causal covariation theory to attribute to HORSE a disjunctive content that not only has as a disjunct *cow-on-a-dark-night* but also *thought-about-hay*.

If we are to hold onto the idea that mind-world causal relations lie at the heart of mental meaning, we need uncover some nonintentional and nonsemantic property that the causal connection between horses and tokenings of HORSE has that that between cows on a dark night and tokenings of HORSE (and thoughts about hay and tokenings of HORSE) does not; a property that explains why the former plays a role in determining the meaning of HORSE whereas the latter does not. The discovery of such a property would facilitate a modification of the causal theory that would solve the disjunction problem.

Fodor considers, and finds wanting, several attempts to solve the disjunction problem in the above described way. Firstly, there is Dretske's (1981) idea that the fundamental, meaning-determining causal connections are those that hold in the period when a symbol is being learnt. According to Dretske, what a symbol means depends upon what property it covaries with, or what information its tokenings carry, in the learning period. Applied to our example the idea would be that HORSE means *horse* because in the learning period horses, and only horses, cause tokenings of HORSE. We get misrepresentation when, outside the learning period - i.e. once the meaning of the symbol has been fixed - a tokening of the symbol is caused by something that doesn't have the property that covaried with the symbol in that crucial period.

Fodor objects<sup>118</sup> that: (i) There is no principled, non-arbitrary way of saying when the learning period ends and misrepresentation becomes a possibility; (ii) The theory can apply only to symbols that are learnt and thus not to the symbols of Mentalese as Mentalese is an innate language; (iii) It is not true of learnt symbols that they covary with the property they express in the period during which they are being learnt. Consider the English word "horse". At best utterances of this word by me were caused by, and only by, horses when I was learning that term. But had I been confronted by a cow on a dark night I would have said "horse". The truth of this counterfactual entails that "horse" doesn't covary with horses in the learning period for the relevant notions of causal connection and causal covariance are counterfactual-supporting.

A second approach to the disjunction problem involves appealing to optimal circumstances. It's true that cows sometimes cause me to think HORSE but, usually, when this happens conditions are not ideal with respect to determining what type of object is before me. This is the case when I am confronted by a cow on a dark night; had conditions been ideal, that is had it not been dark, I would not have mis-identified the cow as a horse. This leads quite naturally to the thought that it is the causal connections that hold in ideal or optimal conditions that determine meaning, where these ideal or optimal conditions are to be described in psychophysical terms, that is, in terms of lighting conditions, spatial relationship to the distal stimuli, and such like. Given this, in cases of misrepresentation a symbol is caused by something that would not have caused it had conditions been ideal. Thus the fact that cows sometimes cause tokenings of HORSE doesn't mean that cows fall within the extension of that symbol (or that it has a disjunctive content one of the disjuncts being *cow*) because cows do not, and would not, cause tokenings of HORSE in optimal conditions.

The main problem with this theory is that it does not apply to belief, or thought in general, (and thus, given RTM, not to the symbols of Mentalese) due to the role of beliefs in belief fixation. Whenever a horse causes me to think that there is a horse present the causal connection between the horse and the thought will be mediated by a

---

<sup>118</sup> See *Psychosemantics* and "Semantics Wisconsin Style".

whole collection of beliefs including beliefs as to what horses look like. A consequence of this is that being in the psychophysically described optimal circumstances with respect to a horse is no guarantee that that horse will cause me to think HORSE. I could be in just those circumstances and have a belief which interferes with this process.

A third attempt at solving the disjunction problem involves appealing to teleological considerations. The basic idea is that a representation expresses the property that it causally covaries with in those circumstances where the mechanism that produces it is performing its proper function. "Proper function" here is to be understood in evolutionary terms so that it is the proper function of mechanism M to produce tokens of representation R in response to instances of property P if and only if M was selected for in virtue of producing R's in response to instances of P. Expressed in terms of the concept of Normalcy, the teleological theory has it that R expresses P if R causally covaries with P in Normal circumstances.<sup>119</sup> Thus in cases of misrepresentation conditions are not Normal; in other words, a mechanism produces a representation in response to a property, and in so doing does not perform its biologically proper function.

Fodor has two compelling criticisms of this theory<sup>120</sup>. One is that representations are produced Normally not only by mechanisms that mediate the causal connections between properties and the representations that express them. In thought, representations are often Normally produced in response to some other representation, or a thought there being no instance of the represented property to be seen. For example, in chains of thought a tokening of HORSE is often

---

<sup>119</sup> Teleological theories of a biological, evolutionary form have been advanced by a number of philosophers in recent years. Some notable examples are: Millikan (1984), (1989), Papineau (1987), (1993), Dretske (1986) and Dennett (1987). Indeed, Fodor briefly flirted with a version of the teleological theory. See his long suppressed paper "Psychosemantics, or: Where Do Truth Conditions Come From". It must be emphasised that the teleological theory described above is the theory that Fodor considers and does not necessarily coincide with all (or any) of the theories just cited. Arguably, however, Fodor's criticisms of the theory he describes cause problems for any teleological theory of a biological, evolutionary hue.

<sup>120</sup> See his 'A theory of Content, I: The Problem'



caused quite Normally by thoughts about hay. This fact would seem to force the advocate of the teleological theory to conclude that HORSE means *horse-or-thought about hay*. In other words the theory doesn't solve the disjunction theory.

Fodor's primary objection to the teleological theory is that it is afflicted with a problem of indeterminacy, and consequently that, once again, it fails to solve the disjunction problem. This can be seen by considering the famous case of the Frog and the fly. Frogs have a mechanism M which, in response to certain distal stimuli, produces a representation R that in turn causes a snap. Flies, when they impinge upon a frog's visual system, typically set off such a causal chain and as a result end up being swallowed and ingested. This is clearly a good thing to happen to a frog, and it is in virtue of such effects that M has been selected for. Now what is the function of M? One possible answer is that it is to detect flies. If M is a fly detector then, according to the teleological theory, R means *fly*. Another possible answer is that, due to the fact that all the flies in the frog's environment are little ambient black things (LABTs for short) and all the LABT's are flies, M is an LABT detector; for M was selected for in virtue of the effects of producing R in response to LABTs. If that is the function of M then, on the teleological theory, R means, not *fly*, but *LABT*. Fodor's crucial point is that there is nothing the advocate of the teleological theory can appeal to in order to justify the preference of one of these stories to any other. In particular counterfactuals cannot be appealed to for the basic idea driving the theory is that it is the mechanism's *actual* history that determines its function. Thus, the teleological theory entails that R has no determinate content. Fodor writes: 'It bears emphasis that Darwin doesn't care which of these ways you tell the teleological story' ('TOCI' p. 72). And again: 'The moral . . . is that . . . Darwin doesn't care how you describe the intentional objects of frog snaps. All that matters for selection is how many flies the frog manages to ingest in consequence of its snapping, and this number comes out exactly the same whether one describes the function of the snap guidance mechanism with respect to a world that is populated by flies that are, de facto, ambient black things, or with respect to a world that is populated with ambient black dots that are, de facto, flies . . . So its no



use looking to Darwin to get you out of the disjunction problem'. (pp. 72-73)

The above described attempts to solve the disjunction problem all involve specifying a property that the causal connection between a representation and the property that it expresses has that the causal connection between the representation and all other properties do not have; a property in virtue of which the meaning-determining causal connections determine meaning. So, according to Dretske's theory, the crucial property is that of being a causal connection that holds in the learning period. According to the second theory it is that of being the connection that holds in ideal or optimal circumstances. And according to the teleological theory it is that of being the connection that holds in Normal circumstances.

All these theories take the causal connections in question to be counterfactual supporting and hence to be causal laws. This has the effect of making them all Type one theories. That is they attempt to define, in naturalistic terms, what Fodor describes as a "Type one situation", a situation that is such that:

(i) If it's a law that Ps cause S-tokens in type one situations, then S means P (and if P is disjunctive then so be it);

and

(ii) not all situations in which S gets tokened qualify as type one, so that tokens of S that happen in *other* sorts of situation s are ipso facto free to be false. ('TOCII' p. 64)

These theories have it that, respectively, the Type one situation is the learning situation, the ideal or optimal situation and the Normal situation.

Fodor's theory points to a property that the causal connection between a representation and the property that it expresses has that marks it apart from all other causal connections involving that representation. But his is not a Type one theory. For he suspects that, because of the robustness of meaning, there just are not any naturalistically describable circumstances in which a representation is

caused by, and only by, instances of the property that it represents.<sup>121</sup> For Fodor, the key difference between the causal connection between a representation and the property that it expresses, on the one hand, and all those causal connections between that representation and other properties, on the other, is that the latter depend on the former but not vice versa. The idea can be brought out by considering the now familiar example of the Mentalese symbol HORSE (which, of course, means *horse*.) Horses cause tokenings of HORSE. Cows (on a dark night) also cause tokenings of HORSE. Now cows wouldn't cause tokenings of HORSE were it not the case that horses did. But not vice versa; for horses would still cause tokenings of HORSE even if cows didn't. In other words, break the causal connection between horses and HORSE and you thereby break the causal connection between cows (and all other non-horses, for that matter); but breaking the latter connection will not thereby break the former. Expressed in Fodor's now famous terminology, the causal connection between cows (and all non-horses) and HORSE asymmetrically depends on that between horses and HORSE. This feature of the causal connections involving HORSE, thinks Fodor, lies at the heart of its meaning. Thus he presents the following as a sufficient condition for a primitive, nonlogical symbol of mentalese S to express a property P:

S expresses the property P if:

- (i) Ps cause Ss; and
- (ii) For any property P\* not equivalent to P, if P\*s cause Ss then the P\* -> S connection asymmetrically depends on the P -> S connection.

---

<sup>121</sup> Fodor says of his theory that it:

Has the desirable property of not assuming that there are such things as Type one situations; in particular, it doesn't assume that there are circumstances - nomologically specifiable and naturalistically and otherwise nonquestion-beggingly specifiable - in which it's semantically necessary that only cows cause "cows". Nor does it assume that there are nonquestion-beggingly specifiable circumstances in which it's semantically necessary that *all* cows would cause "cows". (TOCII' p.91)

Actually the above only counts as a first approximation of Fodor's theory. This is because he conceives the causal connections in question as being counterfactual supporting. He holds that 'if the generalisation that Xs cause Ys is counterfactual supporting, then there is a "covering" law that relates the property of being X to the property of being a cause of Ys: counterfactual supporting generalisations are (either identical to or) backed by causal laws, and laws are relations among properties' ('TOCII' p. 93). Therefore a better expression of his theory would be:

S expresses the property P if:

- (i) It's a law that Ps cause Ss; and
- (ii) For any property P\* not equivalent to P, if its a law that P\*s cause Ss then that law asymmetrically depends on the law that Ps cause Ss.<sup>122</sup>

Applied to HORSE Fodor's idea is that that symbol expresses the property horse, that is, means *horse*, because it's a law that horses cause tokenings of HORSE, a law upon which all causal laws relating non-horses to tokenings of HORSE asymmetrically depend. Thus HORSE doesn't mean *cow on a dark night, horse-or- cow on a dark night* or *thought about hay* despite the etiological heterogeneity of HORSE tokens.<sup>123</sup>

To say that a law, or a causal connection, asymmetrically depends upon another is to commit oneself to the truth of certain counterfactuals; specifically, it is to say that if, contrary to fact, the latter law did not hold, then neither would the former, but not vice

---

<sup>122</sup> Here I am assuming that the sentence "there is a nomic relation between the property of being X and the property of being a cause of Ys" is semantically equivalent to (the somewhat less unwieldy) "it's a law that Xs cause Ys".

<sup>123</sup> Here's how Fodor puts it:

So, what the story about asymmetric dependence comes down to is that "cow" means *cow* if (i) there is a nomic relation between the property of being a cow and the property of being a cause of "cow" tokens; and (ii) if there are nomic relations between other properties and the property of being a cause of "cow" tokens, then the latter nomic relations depend asymmetrically upon the former. (p. 93)

versa (that is, the latter would hold even if, contrary to fact, the former didn't). If, following Lewis (1973) and Stalnaker (1984), we understand counterfactuals in possible world terms, then to say that law Y asymmetrically depends on law X is to say that in nearby (or, perhaps, the nearest) possible worlds where it's not a law that X, neither is it a law that Y; and in nearby (or in the nearest) possible worlds where it isn't a law that Y, it is nevertheless a law that X. Thus, Fodor can be understood as claiming, for example, that HORSE means *horse* if/because in the nearby possible world where it's not a law that horses cause tokenings of HORSE neither is it a law that cows, thoughts about hay, etc. cause tokenings HORSE; but not vice versa. Fodor expresses some qualms about such a possible-worlds understanding of his theory. However, he does talk in possible-worlds terms, and often attempts to justify his theory by pointing out how things are in nearby possible worlds. Following this convention, I will understand the notion of asymmetric dependence in possible world terms. I think there is a very good reason for doing this: namely, it gives us some tangible grasp on what Fodor's theory comes to, thereby putting us in a better position to evaluate it.

### 7.3 Natural language

My first step in evaluating Fodor's theory will involve addressing the question of whether it applies to the primitive nonlogical symbols of natural language. As we shall see, this question is highly relevant despite the fact that Fodor's primary concern is with Mentalese. There is a very close relationship between natural language and thought. For example, it is by means of language that we express our thoughts, so that underlying any causal or nomic connections between properties and the natural language symbols that express them will be causal or nomic connections between those very properties and the corresponding symbols of mentalese. Consequently, if the theory works for natural language then there is good reason to take it seriously as a theory of Mentalese. Indeed Fodor seems to be relying on this consideration as he often appeals to (putative) facts about natural language in motivating and defending his theory. And he often presents natural language examples, implying that what is true of the natural language symbol in

question is just as true of its Mentalese analogue; it's just that English is "easier to spell".<sup>124</sup> Therefore it would be bad news for Fodor if his theory didn't apply to natural language, or only applied to a narrow portion of it. If this were the case serious doubts would arise as to whether there were any reasons for believing that the theory applied to Mentalese; much of the motivation for endorsing the theory would have evaporated. Hence the relevance of addressing the question of whether the theory applies to natural language symbols.

So the question is: does Fodor's theory work for the primitive nonlogical symbols of natural language? Do the primitive nonlogical symbols of English even satisfy Fodor's sufficient condition? If one restricts one's attentions to such words as "horse" and "cow" the

---

<sup>124</sup> Fodor introduces and motivates his asymmetric dependence story in *Psychosemantics* in the following manner. He begins by announcing that he is after 'a difference between A-caused "A" tokenings and B-caused "A" tokenings that can be expressed in terms of nonintentional and nonsemantic properties of causal relations' (p.106). To discover this difference he considers a specific symbol, namely the English word "horse", the tokenings of which are sometimes caused by horses and sometimes by cows (when seeing a cow I misidentify it as a horse). Fodor is explicit that what is true of the English word "horse" will also be true of its mentalese counterpart: 'Here we have all the ingredients of the disjunction problem (set up, as it happens, for a token of English rather than a token of Mentalese; but none of the following turns on that)' (p. 107). A consideration of this case leads him to the conclusion that:

misidentifying a cow as a horse wouldn't have led me to say 'horse' *except that there was independently a semantic relation between 'horse' tokenings and horses*. But for the fact that the word 'horse' expresses the property of *being a horse* (i.e. but for the fact that one calls horses 'horses') it would not have been *that* word that taking a cow to be a horse would have caused me to utter. Whereas, by contrast, since 'horse' does mean *horse*, the fact that horses cause me to say 'horse' does not depend on there being a semantic - or, indeed, any - connection between 'horse' tokenings and cows. (pp. 107-108)

In 'TOCII' Fodor again appeals to a natural language example to introduce and motivate his theory of mental content. Here he says that the linguistic practise of using the expression "fetch me a slab" to request a slab (and hence the causal connection between a desire for a slab and an utterance of "slab") asymmetrically depends on the practice of using "slab" to predicate slabhood (and hence the causal connection between slabs and utterances of "slab"). See pp. 96-100.



answer would seem to be that they do. I often respond to horses by saying, either aloud or to myself, "horse". And if there were doubts as to whether an individual (for example, a foreigner or a child) understood the word "horse", those doubts would be confirmed if she didn't respond to horses by saying "horse"; especially if she tended to respond to horses by uttering some other word and tended to utter "horse" in response to some other type of object. So it would appear that it is true of individuals who mean *horse* by "horse" that there is a reliable causal connection (and perhaps even a nomic connection) between horses and their uttering "horse". Moreover, this connection would appear to be basic with respect to all the causal connections between non-horses and "horse". "Horse" is not a symbol of an innate language, rather it has to be learnt. The learning process involves a teacher trying to establish in the learner a disposition to respond to horses with "horse". Until such a disposition has been established (and hence a reliable causal connection between horses and "horse") the teaching process will not be deemed to have been successful; that is, the language learner will not be judged to have grasped the meaning of "horse". In establishing this disposition, a whole load of other dispositions to respond to non-horses with "horse" will thereby be established; for example, the dispositions to produce "horse" in response to cows on a dark night, pantomime horses, pictures of horses, thoughts about hay, and so on. Thus we have asymmetric dependency of sorts, for: had the causal connection between horses and "horse" not been established none of the other connections would hold; and these latter connections are a contingent by product, or side effect, of the former.<sup>125</sup>

Thus it is arguable that "horse" comes close to satisfying Fodor's condition. But a mere satisfaction of the condition does not entail that the Fodorian theory applies to such natural language symbols as "horse". One possibility is that though Fodor's condition is sufficient it is not in virtue of its satisfaction that "horse" means *horse*. If this

---

<sup>125</sup> In saying that we have a case of asymmetric dependence here I don't mean to imply that the nature of the relationship between the various causal connections is just as Fodor describes. All I mean to say is that there is an intuitively obvious respect in which all the connections between non-horses and "horse" depend upon that between horses and "horse", but not vice versa.

were the case his theory would not tell us in virtue of what "horse" means *horse*. Another possibility is that the condition for meaning that Fodor presents is not sufficient but, rather, necessary. In the case of "horse" I think that at least one of these possibilities is the case.

My intuition is that to say that "horse" (on my lips) means *horse* because of the nature of the causal connections in which "horse" figures (and the dependency relationships between them) is to get things the wrong way round; it is to put the cart before the horse. When, in response to a horse, I say "horse", what I am doing is predicating horsehood of the beast before me. That I do this by uttering "horse" is a product of my understanding of that word, of its meaning for me. If I understood "horse" differently then I would use some other word. Similarly, it is because I understand "horse" to mean *horse* that I utter that word when I mistake a cow on a dark night for a horse. And again, it is my understanding of "horse" that explains why my desire to tell you about Dobbin manifests itself in my uttering "horse". In short, it is meaning that determines causal connections rather than the other way round. The upshot of this is that even if the causal connections being as they are described by Fodor is sufficient for "horse"'s meaning *horse* (in the respect that "horse" couldn't mean anything other than *horse* given that the connections are this way) it is not in virtue of their being this way that that word has that meaning.

A further intuition of mine is that the causal connections being as Fodor describes is a necessary (rather than a sufficient) condition for "horse"'s meaning *horse*. In order to mean *horse* by "horse" it is necessary that I use that word to predicate horsehood of horses. A failure to do this would suggest that I had no idea what a "horse" was. Thus it is necessary that horses causes me to utter "horse". After all we suspect that foreigners and children who don't generally say "horse" in response to horses (who say nothing at all or utter some other word) do not understand or grasp the meaning of the English word "horse". Why would we do that if Fodor's condition wasn't necessary? And so long as he has only a necessary condition, the worry will be that the determinants of the meaning of our mental states are non-naturalistic facts about us and those states.

Is the condition even satisfied by other primitive nonlogical symbols of English? After all such words as "horse" need not be

typical. The now familiar reflections of Burge and Putnam suggest that at least some words are such that their meaning on our lips is partly determined by facts about the world external to us, including facts about the social and linguistic communities to which we belong. What this suggests is that what many of the words in my idiolect mean is determined by what they mean in my linguistic community, given my intentions and desires to mean by my words just what my fellows mean. I am a bona fide member of an English speaking linguistic community. The words of that community have a meaning that I play no role in determining but that meaning can, and often does, play a fundamental role in determining what various words mean on my lips. For example, the word "protein" has a meaning on my lips despite the facts that I have a limited understanding of what proteins are and that I would win no prizes in a protein-spotting competition. What I mean by this word is what everyone else means, namely *protein*. And I mean this because the meaning (on my lips) of many of the words that I use is borrowed or inherited from that of the words of my fellows, given the fact that I intend and desire to mean just what they mean. Thus, with at least some words, it is not a speaker's dispositions to use them that determines meaning; and there will be many cases where a speaker means such and such by a word despite the fact that she does not satisfy Fodor's condition for meaning such and such by that word.

In *The Elm and the Expert* Fodor develops a line of thought with respect to deferential concepts that constitutes a reply to the kind of objection presented in the preceding paragraph. Consider the word "elm", for example. Most people cannot recognise an elm when they see one; they can't tell elms from beeches, for instance. This fact notwithstanding, argues Fodor, there is still a reliable causal connection between elms and "elm", a connection that is mediated by experts. I do not normally care to what species the tree before me belongs. In those cases where I do care I can utilise an expert so that I will respond by uttering "elm" to (and only to) elms. The situation is not appreciably different from that where we employ instruments of observation or laboratory equipment in order to determine whether or not we are being confronted with an instance of a particular property. In other words, just as there is a causal correlation between acidhood and "acid", there is a causal correlation between elmhood

and "elm" and between proteinhood and "protein"; in the first case the connection is mediated by an instrument of observation (i.e. a piece of litmus paper) and in the other two it is mediated by an expert.<sup>126</sup>

I have several objections to this line of thought. First, even if it is true that there is an expert mediated reliable causal connection between proteinhood and my utterances of "protein", it doesn't follow that it is in virtue of the existence of this connection that "protein" means *protein* on my lips. Describing the situation tends to do nothing other than suggest that "protein" means what it does on my lips in virtue of the facts that it means *protein* on the lips of officially recognised experts and that I am willing to defer to these experts and intend to mean by "protein" what they mean by that word. It is also worth pointing out that there is a key difference between expert mediated casual connections, on the one hand, and instrument of observation mediated connections, on the other. Employing an instrument of observation is an active process that involves a subject's utilising a body of knowledge that she has about significant properties of the referent of the term in question. For example, when I determine that the solution is acid by dipping in a piece of litmus paper, I utilise my knowledge that acid turns litmus paper red. Because of this, it seems natural to say that *I* determine the

---

<sup>126</sup> Here is how Fodor makes the point:

"I can't tell elms from beeches, so I defer to the experts." Compare: "I can't tell acids from bases, so I defer to the litmus paper"; or "I can't tell Tuesdays from Wednesdays, so I defer to the calendar." These three ways of putting the case are, I think, equally loopy, and for much the same reason. As a matter of fact, I *can* tell acids from bases: I *use the litmus test to do so*. And I can tell elms from beeches too. The way I do it is, I consult a botanist.

What I do with the litmus, and with the botanist, is this: I construct environments in which their respective states are reliable indicators of the acidity of the sample and the elmicity of the tree; in the one case, I dip the litmus into the fluid, in the other case. I point the expert at the tree... From the point of view of an informational semantics, the situation is *absolutely normal*: that my *elm* and my *acid* thoughts have the content that they do depends on their being mechanisms that reliably correlate them with instantiations of elmhood and acidhood respectively. (*The Elm and the Expert* pp. 34-35)



acidhood (or otherwise) of the solution. Yet when I defer to an expert as to the type of tree before me, I do not engage in an active process that involves utilising my knowledge as to some of the key properties of elms; rather, I passively respond to someone who tells me what to think and say. Because of this fact, it seems wrong to say that I can detect elms and tell elms from beeches. In other words, expert mediated causal connections cannot be lumped together with instrument of observation mediated causal connections.

Second, Fodor runs the risk of making it far too easy for a subject to grasp the meaning of a word or concept. Suppose that Edgar is employed by an ornithologist conducting a survey of the birdlife in a particular stretch of woodland. He tramps round the wood with the bird expert writing down the name of the species of any bird that the expert sees and subsequently calls out. (For example, when the expert sees a Jay he says "Jay" and Edgar, as instructed, writes down "Jay" in his notebook.) Edgar knows nothing about the creatures whose presence he is recording, indeed he doesn't even know that the survey is a survey of birds. For all he cares, it could be a survey of mushrooms or wild flowers. I take it that Edgar does not grasp the meaning of the word "Jay" or the meanings of any of the words that he writes down in his notebook. Yet there is as much of a reliable expert mediated causal connection between Jayhood and Edgar's use of the word "Jay" as there is between proteinhood(/elmhood) and my use of the word "protein"(/"elm"). In short, by appealing to experts in order to deal with the charge that his theory does not apply to many of the words that belong to my idiolect of English, Fodor runs the risk of implying that there is meaning in cases where there clearly isn't any, thus providing a *reductio* of his theory.

Third, suppose that, once more, we accept that there is an expert mediated causal connection between proteinhood and my use of "protein". But it is far from clear that this connection is of the type that Fodor requires. As will become clear later (see section 7.4), the fact that proteins sometimes cause "protein" (i.e. in those cases where I have an expert with me) far from implies that it is a law that proteins cause "protein". Moreover, given that an instance of proteinhood will cause me to utter "protein" in the company of an expert only if I believe that expert to be an expert, it would appear that the most accurate way to describe the nature of the connection is



in terms such as these: proteins cause me to utter "protein" when I am in the company of someone who I believe to be a protein expert. In other words, Fodor's appeal to experts doesn't establish that proteins reliably cause "protein" but only that proteins reliably cause "protein" when the subject is in the company of experts who are believed to be such. But the latter connection is inherently intentional and so is of no use to Fodor in his project of producing a naturalistic theory of content.

Finally, it is far from clear that proteins cause me to utter "protein" even when I have an expert at hand. Fodor would no doubt argue along the following lines: in such cases an instance of proteinhood causes the expert to have a protein thought, a thought that causes him to utter "protein", an utterance that causes the subject to have a protein thought and to utter "protein". Given the transitivity of the causal relation, all this implies that an instance of proteinhood causes the subject to utter "protein". However, it strikes me as mistaken to argue in this manner. Surely what causes the subject to utter "protein" is not the protein but the expert's pronouncement. In other words, we do not have a case where an instance of proteinhood causes the subject to utter "protein". If that sounds implausible consider this case. One of my housemates hears a noise in the back-yard and comes to the conclusion that the house is being burgled. She bursts into my room shouting "we're being burgled" causing me to form the belief that my house is being burgled. What caused that belief? Surely not the noise in the back-yard (a noise that I never even heard). Rather, the cause of my belief is my housemate's pronouncements. I don't see how the expert case is any different: just as the noise in the back yard does not causally impinge upon me, neither does the instance of proteinhood causally impinge upon the subject. Therefore, proteins don't cause me to utter "protein", experts notwithstanding.

Many of the words that we use mean what they mean on our lips (partly) because of the knowledge or the beliefs that we hold concerning the nature of the property that the word in question expresses. I mean *chair* by "chair" partly because of the fact that I believe chairs to be manmade objects with backs and legs that are used, and are designed to be used, as seats for one person. If the beliefs of mine that were associated with the word "chair" were

radically different from what they are then, on my lips, that word would not mean *chair*. Reflecting this, when we want to determine whether an individual understands a word X, or grasps its meaning, we often ask such questions as "what is an X?", "what are X's?". If no answer is forthcoming, or if the answer is radically different from the one that we would give, we conclude that the individual does not mean anything by the word in question, or means something somewhat different from what we mean. So, for example, you can't mean *chair* by "chair" if you don't believe that the items that fall within the extension of "chair" are backed, legged, one person seats. And if you do mean *chair* by "chair", it is partly in virtue of your having these beliefs.

The preceding comments might raise the suspicion that I run the danger of lapsing into some unacceptable version of semantic holism or endorsing something like a description theory of reference. Such worries, though natural, can easily be allayed. To say that one has to hold certain beliefs about the nature of the referent of a word for that word to mean such and such on one's lips is not to imply that for any two individuals to share a meaning they must have associated with their respective words just the same collection of beliefs. To see this consider the following. Within a linguistic community at a particular point in time there is often a collection of widely held, orthodox beliefs concerning the nature of the referents of various of the terms of its language. For example, with respect to the English word "water" the current orthodoxy within our linguistic community is that, *inter alia*, water quenches thirst, can exist in a gaseous, liquid, or solid state, is H<sub>2</sub>O, and so on. Learning the meaning of "water" involves learning what water is, which involves learning all, most, or at least some of these facts about water. An individual who was deemed not to know most, or at least some of these facts, would be judged not to understand or have grasped the meaning of "water". There are many other examples: if you don't know that what we call "snakes" are leg-less reptiles, that what we call "whales" are marine mammals, that what we call "ghosts" are restless spirits of the departed, or that what we call "futons" are Japanese sofa beds, then you do not know, respectively, what the words "snake", "whale", "ghost" and "futon" mean. In short, it is true of many words that to understand them or grasp their meaning one has to buy into the

current orthodoxy concerning the nature of their referents. That is why the questions "what does 'X' mean?" and "what is an X?" (or "what are X's?") are often interchangeable.

But of course orthodoxy changes; the received wisdom as to the nature of water varies from group to group and from time to time. Consequently, what one has to know or believe to know what "water" means (or for "water" on one's lips to mean *water* ) depends upon which social group you are a member of and the point in time at which you exist. So, for example, I have to know that water is H<sub>2</sub>O in order to mean *water* by "water" but Aristotle didn't have to know this fact in order to mean *water* by the ancient Greek analogue of "water". An unacceptable holism that implied that two individuals could mean the same thing by a word only if they had identical beliefs concerning the nature of the referents of that word is thus avoided.

These thoughts could no doubt do with a good deal of elaboration and qualification: to which words do they apply?; how much of the current orthodoxy does one have to accept?; is it possible to reject some, or all, of the current orthodoxy and still mean *water* by "water" - if, say, one is aware of that orthodoxy but finds it wanting in some way?; can children get away with knowing less?; and so on. I am not going to try to answer any of these questions, for I think I have established my point that for natural language it is sometimes the case that the meaning of a word on an individual's lips is at least partly determined by what she believes about the nature of the referent of the word in question. Thus the Fodorian view that meaning is wholly determined by the nature of the causal connections between, or causal laws relating, (instances of) properties and (tokenings of) symbols, and the dependency relationships between these connections or laws, would appear not to apply to natural language.<sup>127</sup>

---

<sup>127</sup> Fodor explicitly says that although beliefs mediate meaning-determining causal/nomic connections, the content of those beliefs plays no role in determining meaning. He refers to the Greeks, who had a word with the same meaning as our word "star" despite the fact that they believed that stars were holes in the sky. As we have seen, such a case doesn't constitute a counterexample to the claim that meaning is partly determined by one's beliefs (and not wholly by causal connections), as what one has to believe to mean such and such by a word is relative

These reflections suggest that there may well be counterexamples in the offing, cases where a word of an individual's idiolect of English satisfies the Fodorian sufficient condition for having a certain meaning (namely the conventional meaning) but doesn't have that meaning due to certain facts about the subject's beliefs concerning the nature of the referents of the word in question. Here is one such counterexample (one that I take to be highly plausible). Edgar has moved to the Scottish Highlands, that last bastion of the pine marten. The pine marten is a small carnivorous, tree climbing mammal, a member of the badger-weasel family. Pine martens are brown with a creamy yellow chest and have a long, bushy tail. They are quick, agile, and wary of humans; but in winter they have a tendency to engage in night time raids on domestic refuse bins in search of food. Many a highlander has woken to find that his refuse-laden bin bags have been decapitated by hungry pine martens.

Before he moved to the Highlands, Edgar had never heard of the pine marten, but he was familiar with a family of small birds known as "martins" (a family that includes the house martin and the sand martin) which are closely related to the swallow. One night Edgar is woken by sounds of scratching and screeching near his bins. When he goes to inspect, he catches a glimpse of a small brown creature as it darts away and discovers that his bin bags have been torn open. The next day he tells his tale to a neighbour who informs him that the intruder was sure to have been a pine marten. Edgar is not told what pine martens are but simply jumps to the conclusion that they are a kind of small bird related to the house martin. So he believes that those creatures known as "pine martens" are birds. Over the

---

to one's position in time, and to which social group one belongs. He also refers to the case of Berkeley who believed that chairs are mental entities yet still meant *chair* by "chair". This case might appear to be a counterexample to what I've said as, one might think, it is part of the received wisdom about chairs that they are physical (and hence not mental) entities. My reply is that Berkeley subscribed to enough of the received wisdom to mean *chair* by "chair". He believed (I presume) that chairs are items of furniture, are one-person seats with backs, and so on. And in believing that chairs are mental entities he didn't take them to be marked out from anything else that we typically take to be a denizen of the physical world. It's not as if he thought that chairs are mental entities but tables not, in which case I might want to argue that he couldn't mean *chair* by "chair".



following few weeks he is regularly disturbed by pine martens; hearing their shrieks and catching fleeting glimpses of them as they scamper away from his bins often causes him to utter "pine marten" (as when he says such things as, "there's a pine marten again"). Thus a reliable causal connection between pine martens and his uttering "pine marten" is established. Sometimes things other than pine martens cause him to say "pine marten", for example, torn bin bags and small brown birds. These latter causal connections ride on the back of that between pine martens and "pine marten", and asymmetrically depend on it. In short, Fodor's condition for Edgar's meaning *pine marten* by "pine marten" is satisfied. However, my intuition is that "pine marten" on Edgar's lips doesn't mean *pine marten*, due to the fact that he thinks that what he calls "pine martens" are small birds. You can be mistaken about what we call "pine martins" and still have succeeded in grasping the meaning of that term (for example you can believe that they feed on insects) but you can't be as mistaken as Edgar. Thus, whatever Edgar means by "pine marten", it isn't what I mean by that term (i.e. *pine marten*), and we therefore have a counterexample to Fodor's theory as a theory about natural language.<sup>128</sup>

---

<sup>128</sup> Fodor might respond to this argument by saying that I haven't produced a genuine counterexample as it isn't a law that pine martens cause Edgar to say "pine marten", as is evidenced by the fact that if Edgar was confronted with a pine marten head on in broad daylight he wouldn't say "pine marten". In response one might make several points. First, the objection runs the risk of committing its advocate to an optimal conditions version of the Type one theory, an option that, as we have seen, Fodor rejects. Second, what's so special about what happens when Edgar is confronted head on by a pine marten? Why does the fact that in such circumstances a pine marten wouldn't cause Edgar to say "pine marten" show that it's not a law that pine martens cause him to say "pine marten", any more than the fact that in certain circumstances a pine marten's impinging on me wouldn't cause me to say "pine marten" doesn't indicate that it's not a law that pine martens cause me to say "pine marten"? Third, an advocate of the objection runs the risk of making it so difficult for there to be a nomic connection between a property and a symbol as to imply that Fodor's condition is never satisfied. I don't want to push such objections as I actually think that there are no laws of the sort that Fodor's theory requires; my point here is merely that if there are any causal laws linking symbols and utterances, it's far from



I therefore conclude that Fodor's theory does not apply to the primitive nonlogical symbols of English. To recapitulate: (i) In those cases where there is a link between meaning and the satisfaction of Fodor's condition the condition would appear to be necessary rather than sufficient. (ii) Sometimes the meaning of a word on an individual's lips is determined by the meaning of that word in the wider linguistic community so that an individual can mean such and such by a word without even satisfying Fodor's condition for meaning such and such. (iii) And the condition isn't sufficient as a word of an individual's idiolect can satisfy the condition for meaning such and such yet not have that meaning because of the individual's beliefs concerning the nature of the referents of the word in question.

Of course the failure of Fodor's theory to apply to natural language does not thereby destroy that theory given that it is a theory about Mentalese. However it doesn't bode well for several reasons.

First, Fodor initially seemed to be presenting his theory as applying to natural language as well as to Mentalese, and to be offering its applicability to the former as evidence for its applicability to the latter.<sup>129</sup> In the face of its failure to work as a theory of natural language, it seems reasonable to ask what motivation we have for believing that it works as a theory of Mentalese. The worry is that we have no motivation.

Second, some of the claims I have made about natural language may also apply to Mentalese. Often what is true of natural language is also true of thought; after all, on anyone's theory there is a very close relationship between language and thought. Consider the fact that

---

obvious that the relationship between the property of being a pine marten and that of being an utterance of "pine marten" by Edgar isn't one of them.

<sup>129</sup> In more recent writings Fodor appears to reject the idea that his theory applies to natural language pointing out that there are several salient differences between such language and Mentalese, (for example, the former, but not the latter, is a public language the use of whose symbols are governed by linguistic conventions and driven by communicative intentions) differences that imply that the nature and origins of mental meaning will be quite unlike the nature and origins of linguistic meaning. This suggests that Fodor's initial appeals to natural language and his use of natural language examples to explicate his theory were merely an expository device. See, for example, Fodor's response to Brian Loar's paper 'Can We Explain Intentionality?' in Loewer and Rey (1991).

the meaning of some words on my lips is determined by the meaning of those words in my linguistic community and not by the causal/nomic connections between my utterances of those words and the properties they express. "Protein" would seem to be an example of one such word. The word "protein" in my idiolect of English does not satisfy Fodor's sufficient condition for meaning *protein*. So, presumably, neither does its Mentalese analogue, that is, that symbol that figures in the thoughts which underlie my "protein" utterances. What this would seem to suggest is that my protein thoughts are protein thoughts (or, in other words, that the mentalese symbol PROTEIN means *protein*) partly because of the meaning of the English word "protein" within my linguistic community.

Next consider the case of those words the meaning of which on an individual's lips would appear to be determined (at least partly) by her beliefs concerning the nature of the referents of the word in question. Surely the thoughts that express these words have their contents partly determined by those associated beliefs. To see this, consider Edgar and the expression "pine marten". Underlying his utterances of "pine marten" will be thoughts in which the Mentalese expression PINE MARTEN figures. It is difficult to see how the satisfaction of Fodor's condition could entail that PINE MARTEN means *pine marten* when, as we have seen, its English analogue has some other meaning (or no meaning at all). For how could his utterances of "pine marten" not mean *pine marten*, if the underlying thoughts that they express had that content? What this case suggests to me is that, to put it paradoxically, Edgar's pine marten thoughts are not pine marten thoughts because he believes pine martens (or what he calls "pine martens") to be birds. Hence the example of Edgar constitutes a counterexample to the Fodorian theory as a theory about Mentalese as much as it does a theory about natural language.

The other claim I made about natural language meaning was that in those cases where Fodor's condition would appear to be satisfied, and where there was some connection between that satisfaction and the meaning of the word in question, the condition is necessary rather than sufficient. I presented "horse" as an example. Once more it is difficult to see how what is true of the English word could not be true of its Mentalese counterpart. If it is a necessary, rather than a sufficient, condition for "horse"s meaning *horse* on my lips that . . .

then surely it is a necessary rather than a sufficient condition for my horse thoughts being about horses (for HORSE's meaning *horse*) that . . . . Therefore, if Fodor's theory fails to apply to the English word "horse" in virtue of providing a necessary rather than a sufficient condition, then it will fail to apply to the Mentalese HORSE for just the same reason.

To conclude this section, the foregoing examination of the applicability of Fodor's theory to natural language was not merely an academic exercise, for it generated substantial objections to his theory of mental content. However, my objections don't end here. The next stage of my argument will focus on Fodor's idea that there are causal laws relating symbols of Mentalese and the properties that they express.

#### 7.4 Laws and content

My tendency to write in terms of causal connections in discussing Fodor's theory might have obscured the important fact that for Fodor it is not so much causal relations between individuals but mind-world causal laws, that is, nomic relations between properties, that determines mental content. Thus it is not merely because individual horses cause me to token HORSE that HORSE means *horse*; rather it is because there is a nomic connection between the property of being a horse and that of being a cause of horse tokens or, in other words, because it is a law that horses cause tokenings of HORSE. Horses could cause tokenings of HORSE as reliably and regularly as you like, but HORSE wouldn't mean *horse* if those causal transactions weren't subsumed by the law that horses cause tokenings HORSE.

This aspect of Fodor's theory is only implicit in *Psychosemantics*, but is stated loud and clear in 'TOC II' where it is utilised to deal with several objections.<sup>130</sup> On the face of it, it seems awfully strong to

---

<sup>130</sup> Amongst these objections are the following: (i) The theory doesn't apply to UNICORN as, given there are no unicorns, unicorns never cause tokenings of UNICORN. Fodor's reply is that the non-existence of unicorns doesn't stop it from being the case that there is a nomic relation between the property of being a unicorn and that of being a cause of UNICORN (or, in other words, from it being a law that unicorns cause UNICORN). (ii) Fodor's theory has the unacceptable consequence of implying that HORSE means not *horse* but *all the horses except Old Paint* because

require that there be such laws. So we might reasonably ask: could it really be the case that whenever a primitive nonlogical symbol of Mentalese expresses a particular property it is a law that instances of that property cause tokenings of that symbol? In this section I will attempt to establish a negative answer to this question; HORSE means *horse* and VIRTUOUS means *virtuous* but that cannot be because it is a law that horses cause HORSE or that virtuous acts cause VIRTUOUS, for there are no such laws.

Just as we can think about things that do not exist, we can think about things that could not, in point of nomological or metaphysical necessity, exist. So, assuming RTM, we can token symbols or expressions of Mentalese that express properties whose instantiation is nomologically or metaphysically impossible. But such properties cannot stand in nomic relations to other properties; thus, if the instantiation of a property X is impossible, then it cannot be a law that Xs cause Y (for any Y). An upshot of this is that Fodor's theory will not apply to any primitive symbol of Mentalese that expressed such a property, and this would be particularly bad news if there were lots of such symbols. Fodor recognises this point, writing that '[it] appears to be quite a strong consequence of the asymmetric dependence story [that]: no primitive symbol can express a property that is necessarily uninstantiated. (There can't, for example, be a primitive symbol that expresses the property of being a round square).' ('TOCII' p. 101).

So the question arises: are there any primitive symbols of Mentalese that express properties whose instantiation is necessarily impossible? I take it that there are not, and could not be any ghosts, so how about GHOST, the Mentalese analogue of the syntactically simple English word "ghost"? Fodor would no doubt reply by saying that GHOST, though syntactically primitive, isn't primitive in the required sense, for it is introduced by definition. One could imagine a mind which sometimes associated syntactically primitive symbols with complex unwieldy definitions of (necessarily uninstantiated) properties so as to facilitate a shorthand expression of such properties. These

---

of the fact that Old paint wouldn't cause HORSE but that horses except Old Paint did, but not vice versa. Fodor's reply is that causal transactions between both Old Paint and horses except Old Paint and HORSE are all subsumed by the one law, namely that horses cause HORSE.



syntactically primitive symbols would not have their meaning determined in the manner that genuinely primitive symbols do; rather they would mean what they mean in virtue of the definitions that they are associated with in long term memory. So, for example, the idea is that GHOST means *ghost* because there is, in long term memory, a sentence of the form "GHOST = . . ." (Where ". . ." is some expression of Mentalese the constituent symbols of which mean what they mean for the standard reasons). It would be a burden for my mind to lug around the complex sentence ". . ." whenever I engaged in thought about ghosts, and so advantageous to employ the shorthand GHOST.

How should we respond to this line of thought? It may well work for syntactically primitive symbols of Mentalese that express such properties as that of being a round square, for such symbols can be introduced by definition (as the symbols that they express can be defined). But how are we to define GHOST? If that symbol is indefinable (as it would appear to be the case) then the suggestion under discussion isn't going to work and the question "why does GHOST mean *ghost*?" is left unanswered. It's no good to say that the indefinability of GHOST is no problem since something like a dictionary definition will do, that is, an entry which doesn't, strictly speaking, define the symbol (or the property that it expresses), but rather specifies its meaning in an imprecise and roundabout way. This would be no help, at least to Fodor, as it would be beset with the kind of problems that he attributes to holist theories; namely, that it has the consequence that two individuals could mean the same thing by GHOST only if their "dictionary definitions" were identical, something that is rarely going to be the case.

What if the other horn of the dilemma is seized and it is argued that GHOST is the sort of symbol that can be defined? The problem with this is that if that symbol can be defined, then so, presumably, can a whole load of others, such as HORSE, for example. And then, we might ask, why couldn't the definition story work for symbols that express properties whose instantiation is possible?; HORSE, for example, springs to mind once more. The theory I am envisaging here is a two-tier theory that suggests that a basic stock of Mentalese terms have their meaning fixed one way (perhaps the way Fodor describes) and that these symbols are then used to construct



definitions of other symbols, definitions that fix the meaning of these latter symbols. The question is what grounds has Fodor got, apart from sentimental attachment, for preferring his atomistic theory - as a theory of the meaning of such symbols as HORSE - to the holist alternative I have just sketched? If GHOST can be defined, why need holism have the dreadful consequences that Fodor describes it as having?

In conclusion, symbols like GHOST would appear to provide Fodor with much more of a problem than he realises, for at best his theory does not apply to them, and he gives no indication as to how they can be dealt with in a manner which does not undermine his theory. Moreover, GHOST is hardly atypical in this respect; there are plenty of properties we are capable of representing in thought whose instantiation is (metaphysically or nomologically) impossible. Examples include the properties of being a fairy, a witch, phlogiston, divine (perhaps), a god (perhaps), and so on: in short, many of the properties appealed to by the holders of mythical, astrological, and (perhaps) religious beliefs, and many of the properties appealed to by the advocates of discredited scientific and quasi scientific theories. Hence this incompleteness of his theory is hardly a minor shortcoming.

My next objection focuses on the problem which Mentalese symbols that express moral and aesthetic properties cause for Fodor. A consequence of Fodor's view that nomic connections lie at the heart of mental meaning is that for two individuals to mean the same thing by a given symbol of Mentalese it must be the case that the basic laws that govern their respective tokenings of that symbol must coincide. For example, suppose that the symbol X belongs to my idiolect of Mentalese and to yours as well. Does my X mean the same as your X? Only if the basic law that governs the tokening of X in me is the very same law that governs the tokening of X in you. If the basic law that governs my tokening of X is "x's cause X" and the basic law that governs yours is "y's cause X", then our respective X's are non-synonymous. What I will try to establish is that this requirement for meaning-equivalence is not satisfied by many of those symbols of Mentalese that express moral and aesthetic properties: in other words, that it is a consequence of Fodor's theory

that VIRTUOUS, for example, diverges in meaning from one person to the next.

Beliefs about horses mediate the causal connection between horses and HORSE. For example, it is because of the beliefs that I have about what horses look like, how they behave and such like, that horses typically cause me to token HORSE when they impinge upon my sensory apparatus. Now most people's beliefs concerning what horses look like, how they behave, and so on pretty much coincide and, consequently, it is no surprise that most of us token HORSE in response to horses rather than, say, in response to chairs. In virtue of this coincidence of horse beliefs, if you and I confront an object and agree on such properties as its shape, size, colour, how it moves, and so on, we will agree on the question of whether or not it is a horse. And because of the truth of our beliefs, we will usually be correct in our judgements. In short, in virtue of sharing a whole bag of true beliefs about horses we will both be subsumed by the law "horses cause HORSE".

Matters seem to be somewhat different with respect to the Mentalese symbol VIRTUOUS (which, of course, expresses the property of being virtuous). There is widespread disagreement concerning virtue amongst those people who do, or are capable of, making judgements as to the moral properties of the acts and individuals that they confront. Some people think that ambition is a virtue, other people don't; some people think that thrift is a virtue, other people don't; some people think that chastity is a virtue, other people don't; and so on. As a result of such differences it is commonplace for two individuals to be confronted with a person or an act and disagree on the question of whether or not they have before them an instance of the property of being virtuous, despite the fact that there is widespread agreement between them as to what other properties the act or person in question has. For example, Edgar and Waldo both hear about a doctor who knowingly administers a lethal injection to a terminally ill patient who has expressed her wish to die. Edgar thinks the doctor is virtuous and applauds his actions, whilst Waldo condemns him and pronounces that he will rot in hell. Such disagreement is unlikely to be isolated, so that Edgar and Waldo will often disagree as to whether specific people and their acts are virtuous. In short, Edgar will often token VIRTUOUS in response

to acts and individuals that Waldo won't, and vice versa. The upshot of this is that it cannot be the case that both Edgar and Waldo are subsumed by the law that virtuous (acts and individuals) cause VIRTUOUS. Thus a consequence of Fodor's theory is that at least one of them doesn't mean *virtuous* by VIRTUOUS; at least one of them will, so to speak, have their Mentalese symbol VIRTUOUS hooked onto the wrong property. And, given that such disagreement is widespread throughout our population, it will be true of many of us that we are *not* subsumed by the law that virtuous (people and acts) cause VIRTUOUS. Thus, a consequence of Fodor's theory is that many of us do not mean virtuous by VIRTUOUS; indeed that many of us do not have thoughts about virtue. I take it that this is an unacceptable consequence; of course we all mean the same thing by VIRTUOUS (otherwise how could our disagreements be disagreements about virtue?). Therefore, Fodor's theory doesn't apply to that symbol.<sup>131</sup>

Parallel arguments will apply to many other symbols of Mentalese that express moral properties, for example, EVIL, SAINTLY, and so on. Put generally, the point is that given the level of moral disagreement in our society, it just isn't the case that all of us are such that our moral concepts are nomically related to the properties that they express. Thus Fodor's theory does not apply to symbols of Mentalese that express moral properties. Neither will it apply, for just the same reasons, to what we might call semi-moral concepts such as HONEST, GENEROUS, KIND, BOASTFUL, VAIN, RUDE, etc.. Moreover, it will not apply to such aesthetic concepts as BEAUTIFUL given the disagreements that people notoriously have as to what is beautiful and what is not.<sup>132</sup>

---

<sup>131</sup> Fodor is quite explicit that his theory is supposed to apply to VIRTUOUS just as much as it is to HORSE. For example, he writes:

All predicates express properties, and all properties are abstract. The semantics of the word "virtuous", for example, is determined by the nomic relation between the property of being a cause of tokens of that word and the property of being virtuous. it isn't interestingly different from the semantics of "horse". (TOC, II' p. 111)

<sup>132</sup> In fact, this argument has still wider application. You and I may well disagree as to what's involved in being intelligent so that you think, for example, that being

It would be no good to respond by arguing that moral and aesthetic concepts are subjective; that, for example, virtue and beauty are in the eye of the beholder. For one thing, such a claim is in tension with Fodor's robustly realist metaphysics. (For the record my intuitions are as realist as Fodor's; of course moral and aesthetic properties are objective). And for another, even subjective concepts have content so we still need an account of where they get their content from. In conclusion then, in virtue of his claim that causal laws lie at the heart of mental meaning, there is a whole battery of symbols of Mentalese that Fodor's theory does not apply to.

None of the above implies that Fodor's theory doesn't apply to such symbols as HORSE. This is because we tend to agree in our applications of HORSE in the sense that it's generally true that if I respond to a perceptual encounter with an object by tokening HORSE, then you would have done too had you been in my shoes. Thus, it may well be the case that it is a law that horses cause HORSE and that we are all subsumed by that law. However, I shall now argue, there is no such law; indeed, for hardly any of the symbols of Mentalese is it a law that instances of the properties that they express cause their tokening.

A first point is that horses could reliably cause tokenings of HORSE without it being a law that horses cause tokenings of HORSE; one possibility would be that whenever a horse caused a tokening of HORSE it was qua having some property other than that of being a

---

good at crosswords is sufficient for being intelligent, an idea that I reject, whereas I think that there is an intimate link between intelligence and verbal articulacy, an idea that you reject. As a result of this difference between us you will token INTELLIGENT when confronted with Waldo the inarticulate crossword ace, whereas I will not. Yet we still both mean *intelligent* by INTELLIGENT.

Another kind of example is represented by such symbols as WITTY and HUMOROUS. Most T.V. comics leave me stony faced yet have a quite different effect on many of my fellows. So I do not token HUMOROUS and WITTY in response to Clive Anderson, French and Saunders, Frank Skinner, and the like. Yet I still mean the same by these symbols as do those folk with very different senses of humour. Such symbols are interestingly different from many of the earlier examples in that the causal connections between individuals and their utterances, on the one hand, and tokenings of HUMOROUS and WITTY, on the other, are not mediated by beliefs or theories as to the nature of wit and humour.



horse. If that were the case, then horse-HORSE causal transactions would be subsumed by some other law. In fact my suspicion is that the many horse-HORSE causal transactions that occur in our world are not all subsumed by the same law; rather, some are subsumed by one law, some by another, some by yet another, and so on. To see this consider the following range of cases. In the first case Edgar, whilst taking a stroll in the country, confronts a horse head on, so to speak. His eyes are wide open, the lighting conditions are favourable and the horse is right before him, filling his visual field. Borrowing a phrase from *Psychosemantics* we might say that Edgar stands in the psychophysically optimal relation with respect to the horse; optimal, that is, for determining such observable properties as its shape, size, colour, and such like. Edgar's visual system duly constructs a representation that explicitly represents these observable properties of the horse, and after this representation has been processed by the object recognition module Edgar tokens HORSE. The question is: was this causal transaction subsumed by the law that horses cause tokenings of HORSE? When addressing questions as to what law is in operation Fodor often utilises the method of differences to settle the issue.<sup>133</sup> Application of the method of differences suggests that in this case it was qua object with a horsy appearance that the horse caused Edgar to token HORSE. *Ceteris paribus*, had Edgar been confronted by anything, be it a horse or not, that didn't have a horsey appearance, then he wouldn't have tokened HORSE. Yet had he been confronted by a non-horse with a horsey appearance (for example, a pair of clowns in a particularly sophisticated pantomime horse costume) he would have tokened HORSE. Therefore, it would appear that the operative law, that is, the law that subsumes this horse-HORSE causal transaction, is not that horses cause tokenings of

---

<sup>133</sup> For example, in 'TOCII' in answering the question of why "horse" doesn't mean *small horse* Fodor writes the following:

As it turns out, routine application of the method of differences suggests that it must be the property of *being a horse* and not the property of *being a small horse* that is connected with the property of *being a cause of "horse" tokens* since many things that have the first property have the third despite their lack of the second: large horses and medium horses simply spring to mind. (p. 102)

See also his treatment of the frog case on pp.106-107.



HORSE but rather that things with a horsey appearance cause tokenings of HORSE.<sup>134</sup>

At the other end of the spectrum there are those cases where Edgar catches a glimpse of a horse out of the corner of his eye and, largely because of the beliefs and expectations he has at that time, this sets off a causal chain that eventuates in a tokening of HORSE.<sup>135</sup> In this kind of case it is not because the creature impinging on Edgar's visual

---

<sup>134</sup> It will do no good to argue that there is a nomic connection between being a horse and having a horsey appearance. Even if there is such a connection it is clearly nomologically possible - as the above example indicates - to have a horsey appearance without being a horse. The application of the method of differences would suggest that it is the property of having a horsey appearance, rather than that of being a horse, that is doing the causal work, just as in the frog case it is the property of being an LABT, rather than that of being a fly, that is doing the causal work, despite the fact that there is a nomic connection between being a fly and being an LABT. (See 'TOCII' pp. 106-107 where Fodor argues that the frog's internal state S means LABT (i.e. *little ambient black thing*) rather than fly, as the operative law is that LABT's cause S, as is indicated by the fact that the frog tokens S in response to LABT's that are not flies, namely bee-bees).

<sup>135</sup> Fodor appeals to this kind of case to deal with the objection that his theory implies that HORSE expresses the property of being a certain pattern of retinal stimulation, namely that pattern that typically mediates horse-HORSE causal transactions. Fodor responds by saying that there is no such typical pattern of retinal stimulation. Sometimes glimpsing a horse out of the corner of one's eye is enough to cause HORSE and in such a case the retinal image is going to be very different from that produced by a head on confrontation with a horse. Here's how he puts the point using the example of COW:

there is no reason at all to suppose . . . that there are specifiable sorts of proximal traces that a cow has to leave on pain of the cow -> COW connection failing. On the contrary, in the usual case there are a heterogeneity of proximal arrays that will eventuate in cow perception, and there's a good reason for this: since, -due to the laws of optics, inter alia- cows are mapped many-one onto their proximal projections, the mechanisms of perception . . . must map the proximal projections many-one onto tokenings of COW. Given the vast number of ways that cows may impinge upon sensory mechanisms, a perceptual system which made COW tokenings intimately dependent upon specific proximal projections wouldn't work as a cow spotter. ('TOCII' p. 109)

apparatus looks to him like a horse that he concludes that there is a horse before him. Rather, such temporary beliefs as that he is in horse infested parts and temporary expectations as that he will soon cross the path of a horse, prime him in such a way that pretty much any object impinging upon him would cause a tokening of HORSE. Consequently, had the horse impinging upon him been a non-horse, a cow, for example, then, *ceteris paribus*, it would still have caused a tokening of HORSE. Thus it is neither *qua* object with a horsey appearance nor *qua* horse that the horse caused a tokening of HORSE; rather it is *qua* impingement on his visual system, or something like that. Hence application of the method of differences would suggest that the operative law is not that horses cause tokenings of HORSE, or that things with a horsey appearance cause HORSE, but rather some other law that is such that pretty much any (medium sized) object capable of impinging on Edgar's visual system satisfies its antecedent.

Between these two extremes will be a whole range of cases of which there is no more reason to think that the subsuming law is that horses cause tokenings of HORSE. Presumably Edgar is not unique; something similar will be true of all of us. Therefore, it would appear that, despite the fact that horses frequently cause us to token HORSE, it is not a law that horses cause tokenings of HORSE.

Could it really be the case that there are laws relating the nonlogical symbols of Mentalese to the properties that they express? Granted, instances of properties often cause tokenings of the symbols that express them, but, for several reasons, the claim that there are such laws has an air of *prima facie* implausibility. The first reason has to do with a fundamental insight of cognitive psychology. Behaviourists such as Skinner held that a subject's behaviour at any point in her history is determined by her current stimulus and her history of reinforcement. The cognitive revolution was partly based on the insight that this just isn't true; how we behave (and also what we think) has an awful lot to do with our current mental states, so making it the case that thought and behaviour is stimulus independent. Now to say that it's a law that, for example, horses cause tokenings of HORSE would seem to imply that when confronted by a horse it is nomologically necessary that I token HORSE in response to it, or, in other words, that I think that there is

a horse present. But if it is nomologically necessary that I think that there is a horse present whenever a horse impinges on my sensory apparatus how could thought be stimulus independent? Stimulus-independence would seem to entail that if there are any mind-world causal laws they will be of the form: Xs cause subjects who are in this, that, and the other intentional state, to token X. Such laws are intentional laws and so can only be expressed using intentional vocabulary. Therefore they cannot be appealed to by Fodor in the course of constructing a naturalistic theory of content. In short, to say that there are laws of the form that Fodor requires would, on the face of it, be to deny the stimulus independence of thought. And surely we don't want to deny that.

A second reason for doubting the plausibility of the claim that there are any of the laws in question is the putative fact that it is commonplace for an instance of a property to impinge on us without our responding by tokening that symbol of Mentalese that expresses that property. I often see horses, grass, and clouds without thinking that there is a horse, some grass, or a cloud before me. It might be objected that I am just not consciously aware of having such thoughts and that in reality, as a result of perceptual interaction with the world, my cognitive apparatus generates a whole battery of representational states (and hence symbols of Mentalese) which never break through to the level of consciousness.

In response to this objection we can recall a point made earlier in the chapter. Each and every object has many many properties. For example, Fang is a dog, is ferocious, is carnivorous, is black, is toothed, etc.. Which properties of Fang are explicitly represented in me when I perceptually interact with him will depend an awful lot on my intentional states, including whatever interests I have and purposes that I am engaged in at the time of interaction. Being a runner with a history of being savaged by large ferocious dogs, my primary aim when I first detect Fang's presence is to determine whether or not he is a large ferocious dog. I am not interested, for example, in the question of whether he's a mongrel or a pedigree. This will have an effect on which properties of Fang are explicitly represented in me, and thus on which symbols of Mentalese are tokened. Given my interests and purposes I token LARGE, FEROCIOUS DOG and not MONGREL. Had I been engaged in a

survey of the number of pedigree dogs in the park then I may well have tokened MONGREL and turned a blind eye, so to speak, to his snarling and other indications of his ferocious nature. What I am saying is that our interests and purposes play a role in determining what cognitive processes are brought to bear on our retinal images and, consequently, on which information implicit in the retinal image comes to be explicitly represented. In other words, when we perceptually interact with an object, which of its properties that we are in a position to detect and represent we actually come explicitly to represent will depend an awful lot on our intentional states. Matters have to be this way because each and every object has so many properties that to represent them all explicitly would be to overburden our cognitive apparatus in every perceptual encounter. (It would overburden our cognitive apparatus for at least two reasons. First, it would take up a lot of space to list all the properties. And second, some properties will require a lot of processing of the perceptual trace and subsequent representations, if their possession by the object of perception is to be discovered. For example, it will take a lot more processing to work out that Fang is a mongrel than it will to work out that Fang is a dog. To do the former will be a waste of valuable cognitive resources if one doesn't care about such questions).

The upshot of these reflections is that it is commonplace for an instance of a property to impinge upon our sensory apparatus without our tokening the symbol of Mentalese that expresses that property. What this suggests, at least on the face of it, is that it is not generally true of the primitive, nonlogical symbols of Mentalese that it is a law that the properties that they express cause their tokening.

In response, it might be said that the laws that Fodor has in mind are *ceteris paribus* laws. According to Fodor, all special science laws are *ceteris paribus* laws; basic laws are to be found only in the domain of fundamental physics. The key point about *ceteris paribus* laws is that they are not exceptionless. Thus it is consistent with its being a law that, *ceteris paribus*, Xs cause Ys, that sometimes Xs occur without causing a Y; in such cases the *ceteris paribus* clause is not satisfied, all else is not equal. However, the failure of an X to cause a Y when all else is equal would constitute a disconfirmation of the claim that it is a law that, *ceteris paribus*, Xs cause Ys. In other words,



ceteris paribus laws are such that the satisfaction of their antecedent guarantees the subsequent satisfaction of their consequent when their ceteris paribus clause is not violated.

However, invoking the notion of a ceteris paribus law will not save Fodor's theory. Reflecting on the nature of such laws indicates that it just isn't the case that, for example, it is a law that, ceteris paribus, horses cause HORSE .

In the course of describing, explaining and predicting phenomena within their domain of enquiry, special scientists make a whole load of assumptions or, as Fodor puts it, operative idealisations, concerning the nature and behaviour of the inhabitants of that domain. Making such assumptions facilitates the discovery and expression of manageable generalisations of considerable explanatory and predictive power. At this point it will be helpful to consider Fodor's favourite example of a ceteris paribus law, a law that comes from the science of geology. The universal, unqualified generalisation that a meandering river erodes its outside bank is clearly false. If the water in a meandering river freezes, or a concrete wall is built along its outside bank, or the tiny abrasive particles in its water are removed, or the world comes to an end, then it will not erode its outside bank. The occurrence of any such event constitutes an unusual, atypical, extra-geological interference with the smooth running of the geological world. When such an event occurs, the operative assumptions or idealisations of geology fail to hold or be satisfied. Geologists do not offer us such false universal generalisations; rather their generalisations are hedged, featuring ceteris paribus clauses. Thus, for example, it is a generalisation of geology that, ceteris paribus, a meandering river erodes its outside bank. To say that, as the geologist does, the statement that, ceteris paribus, a meandering river erodes its outside bank expresses a law is to say 'something like "A meandering river erodes its outside bank in any nomologically possible world where the operative idealizations of geology are satisfied"' (*Psychosemantics* p. 5). In effect then, the ceteris paribus clause is shorthand for something like this: "so long as this, that, and the other event does not occur" (where "this, that and the other event" are events the occurrence of which would result in a failure of the operative idealisations of the science in question to be satisfied).



Clearly there are many distinct event-types the tokening of which would result in a failure of the operative idealisations of geology to be satisfied. Hence it would be difficult, if not impossible, to cash out the *ceteris paribus* clause of the above generalisation. Moreover, the *ceteris paribus* clause could not be cashed out in geological terms given that the interfering events are non-geological. Fodor expresses the point in these terms:

. . . it simply isn't true that we can, even in principle, specify the conditions under which - say - geological generalizations hold *so long as we stick to the vocabulary of geology*. Or, to put it less in the formal mode, the causes of the exceptions to geological generalizations are, quite typically, not themselves, *geological* events. . . . All you can say that's any use is: If the generalization failed to hold, then the operative idealizations must somehow have failed to be satisfied. (*Psychosemantics* p. 6)

Exceptions to the generalizations of a special science are typically *inexplicable* from the point of view of (that is, in the vocabulary of) that science. That's one of the things that makes it a *special* science. But, of course, it may nevertheless be perfectly possible to explain the exception *in the vocabulary of some other science*. In the most familiar case you go 'down' one or more levels and use the vocabulary of a more 'basic' science. (The current failed to run through the circuit because the terminals were oxidised; he no longer recognizes familiar objects because of a cerebral accident. And so forth.) The availability of this strategy is one of the things that the hierarchical arrangements of our sciences buys for us. (*Psychosemantics* p. 6)

We are now in a position to see that there is a significant difference between the law that, *ceteris paribus*, a meandering river erodes its outside bank, on the one hand, and the putative law that, *ceteris paribus*, horses cause HORSE, on the other; a difference that suggests that the latter could not be a law. In the normal course of things the operative idealisations of geology hold. If this were not the case the generalisations presented to us by the geologist would be of such limited explanatory and predictive value that we would have every

right to question the viability of geology as a serious science. When the operative idealisations are not satisfied, when all else isn't equal, generally speaking what has happened is that some extra-geological event has interfered with the smooth running and normal operation of the geological realm. (What has happened is analogous to a mischievous child's opening the oven door whilst a cake is baking inside the oven.)

As we have seen, it is commonplace for an instance of a property to impinge upon an individual without their subsequently tokening the symbol of Mentalese that expresses that property. I often see, hear, and smell horses without tokening HORSE. Usually when I fail to so token HORSE nothing strange, unusual or untoward has happened. The case where, because of my interests and purposes, or because of my occurrent beliefs, I don't notice, so to speak, that the creature before me is a horse, is not one where the operative idealisations of psychology have broken down; if it were then the operative idealisations of psychology would break down in virtually every perceptual encounter - something which would suggest that there was something radically wrong with the psychology that made these idealisations. Moreover, no extra psychological event that interferes with the normal operation of my mind will have occurred, as would be the case were I struck down by lightning before I had the chance to token HORSE, or were I suffering from some neurological disorder. Usually when I fail to token a symbol of Mentalese in response to an instance of the property that it expresses my failure is perfectly explicable in psychological terms; such an explanation would appeal to such intentional phenomena as my beliefs, interests and purposes, and the like. Consequently, the differences between such putative laws as that, *ceteris paribus*, horses cause HORSE and the *bona fide ceteris paribus* laws of the special sciences are such as to suggest that the former are not laws at all. In other words, the laws that Fodor claims lie at the heart of mental meaning are non-existent. This, I take it, is bad news for his theory.

Fodor's claims are implausible largely because of the fact that the causal process leading from an object (or a perceptual encounter with an object) to the formation of a belief or thought as to what is before the subject, is a cognitively penetrable process. However, not all mind-world causal transactions are cognitively penetrable. For

example, many of the sub-personal processes that take place in the early stages of perception are in no way influenced by our beliefs and other such personal level mental states.<sup>136</sup> As such processes involve the production and manipulation of representations, there are, presumably, laws relating external stimuli to the tokening of such sub-personal representations. These representations, unlike the symbols of Mentalese that Fodor focuses his attention on, are not the vehicles of belief content. This raises the possibility that Fodor's theory, though not applying to those symbols that express the contents of our beliefs, does apply to the representations that figure in the sub-personal processing that underlies and facilitates perception and cognition. Given this possibility, even if the above criticisms are conclusive, Fodor's theory still has considerable interest and importance.

### 7.5 Asymmetric dependence

Arguably the most audacious and brilliant aspect of Fodor's theory of content is its invocation of the notion of asymmetric dependence. Asymmetric dependence, he thinks, is the key to solving the disjunction problem, for it allows the informational theorist both to recognise the possibility of, and to give an account of, misrepresentation and representation in thought. In this section I will focus my attention on this aspect of Fodor's theory. I will raise the questions of whether there is as much asymmetric dependence around as he requires, and of whether the appeal to asymmetric dependence implies that our representations have contents quite other than the contents they actually have.

In describing Fodor's theory I argued that with respect to such English words as "horse" the claim that the various causal connections involving those symbols are related as he describes, and

---

<sup>136</sup> Evidence of the fact that some such processes are cognitively penetrable is provided by the fact that optical illusions often take us in when we have knowledge that, one might think, should prevent such illusions from taking place. In *Modularity of Mind* Fodor points out that the Muller-Lyer illusion still works - that is, we cannot but see the parallel lines in the diagram as being of different lengths - even when we know that the lines in the diagram have the same length. (Fodor introduces examples such as this to support his modularity thesis.)

that their being so related is intimately bound up with meaning, has considerable intuitive appeal and plausibility. In teaching me the English word "horse", my teachers attempted to establish a disposition in me to respond to horses by saying "horse"; in other words they attempted to establish a causal connection between horses and my uttering of "horse". They succeeded in doing this, a task that they would have to have succeeded in for me to have grasped the meaning of "horse", but in so doing they inadvertently set up a whole load of other causal connections such as that between cows on a dark night and my uttering of "horse". In an intuitively obvious respect, these latter connections depend on the former in a way in which the former does not depend on them. Moreover, the relationship between these connections would appear to be such that one can break the latter without breaking the former, as Fodor's characterisation of asymmetric dependence demands. This is because of the fact that we have a capacity to discover our mistakes and alter our subsequent behaviour in the light of such discoveries. Suppose that on several occasions I confront a cow on a dark night and jump to the conclusion that the creature before me is a horse. On hearing the cow "moo" I discover my mistake. I then decide not to be so rash in future and reserve judgement as to the nature of any creature that I meet on a dark night. My making this resolution results in a breakdown of the connection between cows (on a dark night) and "horse", whilst leaving the horse-"horse" connection very much intact. Therefore, in the nearby possible world where it's not a law that cows (on a dark night) cause "horse", it is a law that horses cause "horse".

However, matters would appear to be different with respect to representations that are produced by processes that are cognitively impenetrable. I can make decisions and endorse beliefs that influence my tokening of such English words as "horse" and those representations that are the vehicles of my belief contents. Thus the causal connections that such symbols bear to external objects and properties is not fixed once and for all. That is the reason why the cow-"horse" connection can be broken without breaking the horse-"horse" connection. But the same is not true of many of the symbols that are manipulated at the sub personal level, for example the symbols employed by our perceptual modules. To see this, consider



the following example. One of the tasks of the visual system is to determine the colour of objects impinging on the subject's visual system. Typically, implicit in the retinal image is information concerning the colour of the object(s) currently impinging on the subject's visual system. This information is extracted and made explicit by processes in early vision. In order to extract this information, the visual system must make all sorts of assumptions concerning how various colours affect the visual system or show up in the retinal image. These assumptions are not explicitly represented but, rather, hard wired into the visual system. A consequence of the making of these assumptions is that, for example, red objects usually cause a tokening of RED (the symbol of the language utilised by the visual module that is the analogue of the English word "red") when they impinge upon a human subject's visual system. Thus it is a law that red causes RED. However, these assumptions do not always hold true; they only apply within a certain range of lighting conditions. Consequently, in certain abnormal lighting conditions it is orange objects that cause RED; there is nothing that the subject can do to stop such cases of misrepresentation or sensory illusion from occurring in such lighting conditions as the processes that generate RED from retinal images are cognitively impenetrable. Thus it is a law that orange causes RED (in these lighting conditions) just as much as it is a law that red causes RED (in normal lighting conditions).

In this case it is difficult to see how the orange-RED connection can be broken without breaking the red-RED connection; it would appear that these connections stand and fall together. Given the cognitive impenetrability of the processes of early vision, the orange-RED connection cannot be broken in a manner analogous to that in which the cow (on a dark night)-"horse" connection can.<sup>137</sup> Given the nature of the human visual system and the laws of optics that hold in our world, orange objects cannot but cause RED (in abnormal lighting conditions) if red objects cause RED (in normal lighting

---

<sup>137</sup> Perhaps I can successfully decide not to form beliefs about the colour of what is before me unless I can be sure that lighting conditions are normal. But orange objects are still going to look red to me in abnormal lighting conditions, a fact that indicates that RED has been tokened even if I have succeeded in resisting the promptings of my visual system.



conditions). To break the former connection or law without breaking the latter would require effecting either substantial changes in the workings of the human visual system or substantial changes in the laws of optics. This would tend to suggest that the possible world in which there is a red-RED connection but no orange-RED connection is at some distance from our world. It is far from obvious that this world is nearer than that in which the orange-RED connection is broken by effecting changes that are such as to bring down the red-RED connection as well. Parallel arguments will apply to all symbols that are manipulated at the sub-personal level by cognitive modules whose processing behaviour is cognitively impenetrable. Thus the appeal to asymmetric dependence doesn't solve the disjunction problem with respect to such symbols; Fodor's theory would seem to imply that RED means *red or orange*. Earlier I raised the possibility that Fodor's theory could apply to such symbols despite its failure to apply to those that express the contents of our beliefs. If the preceding reflections are anything to go by, this hope would seem to have been somewhat premature. This is a significant result, for a failure of Fodor's theory to apply to those representations that figure prominently in cognitive psychological theory and explanation would constitute an important limitation of that theory. The worry is that Fodor's fixation with the familiar propositional attitudes of folk psychology has resulted in him producing a theory which has an air of *prima facie* plausibility if we concentrate our attentions on those states, but that looks increasingly unsustainable when we reflect upon many of the representational states that are attributed to us by cognitive psychologists.

Another objection to the asymmetric dependence aspect of Fodor's theory is based upon a general idea concerning the relationships that laws bear to one another. Fodor writes as if it makes sense to talk about nearby worlds whose laws are just like ours apart from with respect to the odd law. But wouldn't a world which was slightly different from ours with respect to its laws have to be a world that is very different from ours with respect to its laws? The laws that hold in our world constitute a complex, interrelated, and finely balanced network. Higher level laws are implemented or underpinned by lower level laws so that to break a higher level law would require all sorts of lower level adjustments, adjustments which would probably

have significant ramifications with respect to other higher level laws, not to mention laws at their own level. Consider the law that water boils at one hundred degrees Celsius. Suppose God wanted to break this law and make it the case that water boiled at ninety degrees Celsius. He couldn't make this *prima facie* tiny adjustment without adjusting a whole load of lower level physico-chemical laws governing the behaviour of molecules and, in particular, their response to heat. Making these adjustments might involve making lots more adjustments at the level of atoms, and in turn, at the level of sub-atomic particles and so on. And making these lower level adjustments would surely have higher level consequences quite apart from that with respect to the boiling point of water; for example, consequences with respect to the boiling and freezing points of other substances. In short, there is no nearby world where the laws are just like the laws in our world apart from the fact that water boils at ninety degrees Celsius; for there would have to be a huge nomological mismatch between our world and the world where water boils at ninety degrees Celsius. At this point one is reminded of chaos theory and the much quoted tale about the ramifications of butterfly wing flappings. Consequently, Fodor cannot legitimately claim that there are laws governing the tokening of mental representations that are such that they can be broken without there being any further wholesale ramifications; who is to say what the consequences would be of breaking the law that cows (on a dark night) cause HORSE? Thus, it seems reasonable to doubt that there are any laws governing the tokening of mental representations that can legitimately be said to depend asymmetrically on any other law in the way that the relationship of asymmetric dependence is understood by Fodor.

## 7.6 Some problem cases

In this section I will examine a number of problem cases, cases where it would appear that Fodor's theory entails that a representation has a content at variance with the one it actually has. The first such case has to do with properties that are, in point of nomological necessity, coextensive. The properties of being water and having a boiling point of one-hundred degrees Celsius are coextensive in this respect.

WATER means *water* and not *stuff that boils at one-hundred degrees Celsius*. The problem is that Fodor's theory would appear to imply that the attribution of the latter content is just as defensible as the attribution of the former. For if it is a law that water causes WATER and that all nomic connections between non-water and WATER asymmetrically depend on this law then it will also be true that it is a law that stuff that boils at one hundred degrees Celsius causes WATER and that all nomic connections between stuff that doesn't boil at one hundred degrees Celsius and WATER will asymmetrically depend on that law. In other words, Fodor is faced with a problem of indeterminacy for his theory doesn't have the resources to justify the attribution of either one of the competing contents *water* and *stuff that boils at one hundred degrees Celsius* in preference to the other. Thus his theory falls victim to the same problem that he argued afflicted, and ultimately sunk, the teleological theory.

There are several ways in which Fodor might attempt to respond to this problem. (i) An individual cannot think that the stuff before her boils at one hundred degrees Celsius unless she has a whole bag of concepts, for example the concepts of a boiling point, of the number one hundred, and so on. Having these concepts is a matter of having in one's idiolect of Mentalese a distinct symbol which expresses each of the relevant properties. Now WATER does not mean *stuff that boils at one hundred degrees Celsius* (or, alternatively, to token the Mentalese sentence THAT'S WATER is not to think that what is before one is *stuff that boils at one hundred degrees Celsius*) because, being syntactically primitive, WATER does not have the required internal lexico-syntactic structure.

In reply, we might ask why being syntactically primitive debars WATER from meaning *stuff that boils at one hundred degrees Celsius*? The fact that that content can be expressed in English only by means of a syntactically complex expression doesn't entail that a syntactically simple symbol of some other language could not have that content. After all, Fodor attributes the content *little ambient black thing* to the Frog's state S, a representational state that has no internal lexico-syntactic structure. Moreover, he presumably wouldn't want to say that the Frog has the concepts LITTLE, AMBIENT, BLACK and THING. It may well be true there is a

syntactic difference of the kind the objection alludes to between our *water* thoughts and our *stuff that boils at one hundred degree* thoughts, but Fodor could not legitimately appeal to this fact to justify the attribution of the content *water* to WATER, for to do that would be to endorse the holist idea that the content of a symbol is partly determined by its relations to other symbols in the language to which it belongs.

(ii) A second attempt to deal with the problem of indeterminacy involves appealing to counterfactuals and runs as follows. Its being a law that stuff that boils at one-hundred degrees Celsius causes WATER is a product of the fact that it is a law that water, and only water, has that boiling point, and that it is a law that water causes WATER. Were it not the case that water had that boiling point then stuff that boils at one-hundred degrees Celsius would not cause tokenings of WATER. Thus, it is possible to break the stuff that boils at one hundred degrees Celsius  $\rightarrow$  WATER connection without thereby breaking the water  $\rightarrow$  WATER connection but not vice versa; in other words, the former connection asymmetrically depends on the latter entailing that, on Fodor's theory, WATER means *water*. Similarly, what all this indicates is that it is qua water, rather than qua stuff that boils at one hundred degrees Celsius, that water causes WATER.

The obvious reply to this is that the possible world where the properties of being water, on the one hand, and of being stuff that boils at one-hundred degrees Celsius, on the other, are separated is hardly going to be a nearby one given that it is no accident that water, and only water, has that boiling point in our world. Who is to say what would cause WATER in such a distant world? Maybe creatures like us couldn't survive in such a world. It is a distinct possibility that in every nearby possible world in which that stuff that boils at one hundred degrees Celsius doesn't cause WATER, water doesn't cause WATER either. This case is quite unlike that of the frog and its typically fly-caused internal state S. In the frog's world all LABT's are flies (and vice versa); in other words the properties of being a fly and that of being an LABT are coextensive. But there are nearby worlds where these properties are separated, for example in the psychologist's laboratory where the LABT's are bee-bees. Thus, there are some grounds for saying that it is qua LABT, rather than qua fly,



that flies cause S as, in the laboratory, frogs will snap at bee-bees until the cows come home. But there are no parallel grounds for saying that it is qua water that water causes WATER.

We now come to the second problem case. Several philosophers have argued that Putnam style twin cases pose substantial problems for Fodor (for example, Baker (1989) and Boghossian (1991)) and it would, perhaps, be surprising if we could complete a discussion of the origins of content without mentioning Twin Earth and XYZ at some point. Here is my version of the worry that Fodor can't handle twin cases. XYZ does not fall in the extension of our water thoughts. However, Fodor's theory would seem to imply that it does. WATER also applies to XYZ. XYZ and H<sub>2</sub>O agree in their superficial properties, in how they look, taste, smell and feel to the human visual apparatus. Consequently, given that H<sub>2</sub>O causes me to token WATER via how it looks, tastes, smells and feels to me, XYZ would cause me to token WATER were I to interact with any. Therefore, if it is a law that H<sub>2</sub>O causes WATER, then it is also a law that XYZ causes WATER. This latter law or connection does not asymmetrically depend on the former for it is not true that it can be broken without breaking the former. To see this consider the following. XYZ affects my twin in the way it does (i.e. in just the same way that H<sub>2</sub>O affects me), because of its microphysical structure, because of the laws that hold on Twin Earth (for example, the optical laws), and because of the nature of my twins psychology, in particular, his sensory apparatus. Consequently, given that the laws of optics etc. are here on Earth just as they are on Twin Earth, and given that my sensory apparatus is just like that of my twin, it is true that XYZ would cause me to token WATER were I to come across any. There are three distinct ways of breaking the XYZ connection. The first involves altering the microphysical structure of XYZ so that it looked, tasted, smelt, and felt to me other than it does, so to speak.; that is, so that it affected my sensory apparatus differently than water does. The second way involves altering my sensory apparatus in such a way that XYZ comes to appear to me differently than H<sub>2</sub>O does. And the third way involves altering the laws of optics, for example, so bringing about a change in how XYZ affects me. The crucial point is that none of these ways of breaking the XYZ ->WATER connection are legitimate; none of the putative possible worlds so generated are

relevant to the question of the dependency relations between the laws under consideration. The world in which XYZ has a different microphysical structure is not a possible world at all given that XYZ's microphysical structure is part of its essence. And when breaking connections between properties and symbols, there are certain features of the subject and its world that have to be held fixed, amongst them the basic nature and functioning of the subject's sensory apparatus and the laws - such as the laws of optics - that underpin all interactions between the subject and the objects that inhabit her world. Consequently, there are no nearby possible worlds in which H<sub>2</sub>O causes WATER but XYZ does/would not; given the nature of XYZ and of my visual apparatus, XYZ cannot but affect my sensory apparatus in just the way that H<sub>2</sub>O does in any world in which the laws of optics etc. are as they are in our world. Therefore, Fodor's theory would seem to entail that XYZ falls within the extension of WATER. But XYZ clearly doesn't fall within the extension of WATER.

There are a couple of objections to this argument. Firstly, there is nothing in the way that Putnam describes the case that rules out the possibility of our coming to be able reliably to distinguish between the two kinds of stuff. If the two stuffs are chemically different, so the thought continues, then surely it will be nomologically possible for us to develop a reliable test (maybe requiring the use of a laboratory and sophisticated apparatus) for telling H<sub>2</sub>O apart from XYZ. This generates the argument that in the nearby possible world where we developed such a test there will be an H<sub>2</sub>O->WATER connection but no XYZ->WATER connection. This world is a legitimate world, relevant to determining the dependency relations between the connections in question for in it our sensory apparatus is just as it is in the real world the laws of optics are just as they are in the real world, etc. .

This reply is unconvincing. In the possible world described, XYZ, in direct confrontations with me (that is, outside of the laboratory setting and without the use of any sophisticated apparatus) will still affect or appear to me in just the same way that H<sub>2</sub>O does. Thus, if it is true that H<sub>2</sub>O would cause me to token WATER in such direct interactions then there will still be an XYZ->WATER connection. For there to be no such connection we would have to stop applying

WATER on the basis of direct interactions with liquids. Of course we could stop doing this. We could, for example, recognising that appearances can be misleading, decide not to jump to any conclusions as to the nature of the stuff before us when that stuff is a colourless, odourless, tasteless liquid; we could decide to make such judgements only in the laboratory setting. Making this decision would have the result of breaking the XYZ->WATER connection whilst leaving the H<sub>2</sub>O connection intact. However, this possibility does not vindicate Fodor's position. Water is very important to us and we are forever coming to the conclusion that the stuff before us is water solely on the basis of a direct perceptual encounter that is not mediated by the use of any scientific apparatus or the administering of any test. Rarely do we employ anything more than our unaided senses. Thus the H<sub>2</sub>O->WATER connection is very much unlike the acid-> ACID connection as the latter is mediated by the use of such things as pieces of litmus paper. Now in the possible world under consideration, the H<sub>2</sub>O->WATER connection is unlike the H<sub>2</sub>O->WATER connection that holds in the real world in that it is mediated by the employment of scientific apparatus and the administering of tests; the inhabitants of that world never think THAT'S WATER solely on the basis of a direct perceptual encounter with water (if they did it would be true of them that XYZ would cause them to token WATER). And this fact entails that their world is not nearby, or at least not relevant, for although the H<sub>2</sub>O->WATER connection is held intact it is not the same H<sub>2</sub>O ->WATER connection that holds in the real world; for example, it is not implemented by the same lower level laws. Intuitively, what Fodor needs is a world where H<sub>2</sub>O causes WATER in just the way that it actually does. In the world under consideration H<sub>2</sub>O causes WATER in a radically different way; to employ Fodor's terminology, all else is not equal in that world. In short, so long as the H<sub>2</sub>O->WATER connection as it is in the real world is held intact, there will be an XYZ->WATER connection.

A second response to the accusation that Fodor's theory cannot deal with XYZ is developed by him in 'TOCII'. This response constitutes a modification of his theory. Here is how it goes. There is no XYZ around here; the property of being XYZ is uninstantiated. For this reason XYZ doesn't fall in the extension of WATER, for to so fall in

the extension of that symbol, XYZ would have to feature in the local environment or in the actual history of tokenings of WATER. Thus Fodor amends his theory saying that for a symbol "X" to mean X it has to be the case that 'Some "X"s are actually caused by Xs' (p. 121).

Quite apart from worries about the resultant capacity of Fodor's theory to deal with symbols that express uninstantiated properties (UNICORN, GHOST, etc.), there are major problems associated with the addition of this clause. Fodor is echoing Putnam's thought that we mean "water" to apply to stuff with the same microphysical structure as the local samples of "water". As a consequence of this fact about us, thinks Putnam, "water" doesn't apply to XYZ given that what we normally apply "water" to has a quite different microphysical structure to XYZ. This is a plausible thought and one that Putnam can legitimately hold. But Fodor would be on shaky grounds if he made a parallel move with respect to the Mentalese symbol WATER. Fodor is assuming that WATER is a kind concept. What right has he got to make this assumption? Nothing about WATER's nomic relations to externally instantiated properties suggests that it is a kind concept any more than a concept that is satisfied by anything that has a certain range of superficial properties; why, for example, doesn't WATER mean something like *colourless, odourless liquid*? If WATER had this latter meaning then XYZ would fall in its extension even if it were the case that for a symbol to express a particular property instances of that property must have caused tokenings of the symbol in question. I have never tokened TREE in response to a Giant Redwood as I have never come across a member of that species of tree. But Redwoods still fall within the extension of TREE as that symbol means *tree*; and given that trees have actually caused me to token TREE, to make that content-attribution is not to reject Fodor's additional clause.

In short then, Fodor's additional clause will only help him avoid the undesirable conclusion that XYZ falls within the extension of WATER if he can justify his assumption that WATER is a kind concept. But I don't see how he can do that by appeal to mind-world causal laws alone. Of course WATER is a kind concept, but presumably that is because of our beliefs, intentions, and the like: for example because it is my intention to think such thoughts as THAT'S WATER only in response to samples of stuff with the same



microphysical structure as the stuff that I normally think THAT'S WATER in response to; because I believe that everything to which WATER applies has the same microphysical structure; because I believe that I would be mistaken if I were to think THAT'S WATER of any sample of stuff that wasn't H<sub>2</sub>O; and so on. Therefore, I think it is safe to conclude that XYZ poses a real problem for Fodor's theory.

A third problem case takes us back into the realm of representations that are manipulated at the subpersonal level by processes that are cognitively impenetrable. One objection to Fodor's theory is that it entails that the symbols of Mentalese do not express properties that are instantiated in the extra mental world but, rather, such properties as that of being a certain sort of proximal stimulus. The idea is that, for example, because whenever a horse causes a tokening of HORSE it does so by causing a token of a certain type of proximal trace, the horse → HORSE connection asymmetrically depends on that between the proximal trace type in question and HORSE.

As we have seen, Fodor's response to this objection is to say, quite plausibly, that there is no single, or small range of, proximal trace types that mediate the horse → HORSE connection. However, this line of response will not work for all mental representations. Recall the symbol RED (not that symbol that features in sentences of Mentalese that express the contents of our *red* beliefs, but that symbol employed in early visual processing to indicate that a red object is impinging upon the subject's visual apparatus). RED covaries quite reliably with a certain pattern of retinal stimulation, namely that pattern that is standardly caused by red objects. That is, red objects typically cause retinal images with a certain intrinsic property, and retinal images with that property invariably cause a tokening of RED. If this were not the case then the visual system would not be able to perform the routine task of extracting information concerning the colour of the object(s) before the subject from the retinal image. Now of course RED expresses the property of being red, but Fodor's theory would seem to suggest that it expresses the property of being a retinal image with a certain type of intrinsic property. There will be many other symbols manipulated at the subpersonal level by cognitively impenetrable processes that cause similar problems for Fodor's theory; RED is not an isolated case that can be brushed under the carpet. It will do no good to try to deal with

this problem by stipulating that the connections relevant to the determination of meaning are those between properties that are instantiated in the extra cranial world and the symbols in question. This is because some such symbols express properties that are instantiated in the head. For example, the primal sketch does not represent the external world as being a certain way, but rather represents certain features of the retinal image from which it was generated.

A final problem case is that of states or events that, though they satisfy Fodor's sufficient condition, do not have any meaning or content. Such cases would appear to suggest that Fodor's condition is not sufficient after all. There are plenty of cases where a law of the form "As cause Cs" asymmetrically depends on a law of the form "Bs cause Cs" where the Cs in question mean nothing at all. Indeed, if Fodor's condition was sufficient, given that causal chains give rise to a species of asymmetric dependence and that every event belongs to a causal chain, every event would mean something.<sup>138</sup>

In attempting to deal with this problem Fodor amends his theory somewhat. He asserts that asymmetric dependence alone is not enough for content, stipulating that there must be robustness as well; thus it is asymmetric dependence plus robustness that is sufficient for meaning. This rules out most Cs from meaning anything as Bs always figure in their etiology; they are never caused by non-Bs and thus are not robust.

One can see the appeal of this response; it is not just an arbitrary stipulation whose only recommendation is that it gets Fodor out of a sticky situation. On the contrary, robustness would appear to be intimately bound up with meaning, for all the clear cut cases of meaningful items (for example, public language utterances and thoughts) would appear to be robust. One of the most important and valuable features of language and thought is that it enables us to represent things (for example, to talk and think about things) in their

---

<sup>138</sup> Causal chains give rise to a species of asymmetric dependence for this reason: Every causal chain leading from A to C will be mediated by an event B. In order to cause C, A has to set off a causal chain that leads to B, an event which in turn sets off a causal chain that leads to C. Consequently, if B didn't cause C then neither would A have, but given that the B->C chain doesn't run through A, B would still cause C even if A didn't.

absence. I don't have to have Fodor or one of his books impinging on me in order to talk or think about him and his writings. However, there is a good reason to reject Fodor's addition of the robustness clause.

There are symbols that are not robust, yet that are caused only by instances of the property that they represent. Transducers are mechanisms that play a fundamental role in mediating the interaction between the mind and the world. There are two basic sorts of transducer. The first sort takes physical, non-symbolic events as input, and produces symbols as output, symbols which are then processed by some cognitive module. The second sort takes mental symbols as input and produces physical, nonsymbolic events as output (for example, neural firings that subsequently cause muscle contractions). My interest is in the first sort of transducer. Such transducers are constructed in such a way as to be sensitive only to certain of the intrinsic physical properties of their input; for example, the transducers that produce the retinal image or grey coding are sensitive only to such properties of the light waves hitting the retina as their intensity, wavelength, and the like. Consequently, the symbolic output produced by a transducer will be wholly determined by the intrinsic physical properties of its input, and each of the distinct symbols of the transducer's language will covary with a specific type of physical input. If, on two distinct occasions, my retina is subject to the same stimulation then, all else equal, identical retinal images will be produced. Now the crucial point is that the symbols produced by such transducers represent or express the very properties that they are sensitive to; for example, the symbols produced by the retinal transducer represent the light intensity values at various points on the retina. Given that transducers produce symbols that represent intrinsic physical properties of their causes and that they are sensitive only to such properties, a transducer will produce that symbol that expresses the property P only in response to events which have P. In other words, transducer-produced symbols are not robust, and thereby constitute a counterexample to Fodor's amended theory. Once again it is symbols produced at the sub personal level that cause him problems.

It might be thought that Fodor can save his theory by appealing to the phenomenon of malfunction. In cases of malfunction,

transducers sometimes produce a tokening of the symbol S that expresses the property P in response to a non-P. It might be thought that this entails that S is robust. However, it does not. The laws governing the production of transducer symbols are *ceteris paribus* laws, and in cases of malfunction the *ceteris paribus* clause of such laws will have been violated. If it is a *ceteris paribus* law that transducers of type T produce symbols of type S in response to inputs with property P then there will be no other laws of the form "Ts produce S in response to inputs that have P'" (where P' is some property other than P). I take it that there would have to be laws of this latter form for S to be robust. Therefore S will not be robust, the possibility of malfunction notwithstanding.

### **7.7 Is there more to meaning than reference?**

Fodor's theory is a theory of reference in the respect that it accounts for, at most, the reference of the primitive nonlogical symbols of Mentalese, and consequently of our thoughts. But, one might object, there is more to meaning than reference, thus entailing that Fodor's theory is at best incomplete. For example, the symbols "water" and "H<sub>2</sub>O" have the same reference but are not equivalent in meaning and the thought that I express with the sentence "water is wet" differs in content from that that I express with the sentence "H<sub>2</sub>O is wet"; these symbols and thoughts diverge in respect of a second component of meaning, namely sense, or the mode of presentation of reference. Familiar facts that are widely taken to establish that there is this second component of meaning are the following: the sentence "water is H<sub>2</sub>O" is informative whereas "water is water" is not; one can rationally adopt different epistemic attitudes with respect to thoughts that have the same referential content or truth conditions. For example, one can sensibly believe that water is wet whilst doubting that H<sub>2</sub>O is wet.

In response to the charge that his theory ignores sense, and for that reason is at best incomplete (in naturalising only one component of content) and at worst flat false (in implying that, for example, "water" and "H<sub>2</sub>O" -and their Mentalese analogues - are equivalent in meaning) Fodor takes the bull by the horns denying that there is such



a thing as sense.<sup>139</sup> He attempts to deal with the above described sense-motivating facts - what he calls the Frege cases - by appealing to syntax. In effect, he wheels in syntax to do the job that sense is widely invoked to do. This is how his idea goes. Propositional attitudes are individuated not just in terms of their mode and content but also in terms of the syntactic properties of their vehicle, so that for you and I to share a belief we must each have a sentence in our respective belief boxes that agrees in both its semantic and syntactic properties. If, for example, the sentence in your belief box is WATER IS WET and that in mine is H<sub>2</sub>O IS WET then our respective beliefs do not belong to the same type, despite the fact that they agree in content.<sup>140</sup> Reflecting this fact, sentences that ascribe propositional attitudes to individuals specify not just the mode and content of the attitude, but also the syntactic properties of its vehicle, that is, the syntactic properties of the sentence of Mentalese that expresses the content of the attitude in question. Consequently, an individual can rationally believe that water is wet whilst doubting that H<sub>2</sub>O is wet, for to do that is not the same as to believe that water is wet and simultaneously doubt that water is wet. And the sentence "water is H<sub>2</sub>O" is informative in a way in which "water is water" is not, for an individual who already has the concept WATER can acquire a new

---

<sup>139</sup> In 'TOCII', in answer to the question 'why doesn't "water" mean the same as "H<sub>2</sub>O"?' he says that these two symbols have the same meaning but goes on to argue that having the concept WATER is not the same mental state as having the concept H<sub>2</sub>O in virtue of the fact that the Mentalese symbol WATER has different syntactic properties from the Mentalese symbol H<sub>2</sub>O. His rejection of the notion of sense is most explicit in 'Substitution Arguments and the Individuation of Beliefs', a paper that opens with the sentence, 'The older I get, the more inclined I am to think that there is nothing at all to meaning except denotation; for example, that there is nothing to the meaning of a name except its bearer and nothing to the meaning of a predicate except the property that it expresses'.

<sup>140</sup> In 'Substitution Arguments and the Individuation of Beliefs' Fodor argues that there is a fourth dimension to propositional attitude individuation, namely causal role. So, for you and I to share a belief we must each have in our belief box a sentence with the same content, the same syntactic properties, and the same causal role. However, by *The Elm and the Expert* Fodor had abandoned this fourth feature of propositional attitude individuation.

belief on being presented with the former sentence but not on being presented with the latter.

If this appeal to syntax works then - given that Fodor has no problems accounting for how some of our internal states can have syntactic properties - so long as Fodor's theory of content works as a theory of reference the problems generated by the Frege cases evaporate. However, there are several reasons for scepticism.

(i) Fodor would appear to be making a claim about folk psychology, namely that it individuates propositional attitudes partly in terms of the syntactic properties of mental representations. Moreover, he seems to be committing himself to the view that ordinary folk, when they ascribe propositional attitudes to their fellows, stick their neck out as to the syntactic properties of some mental representation. Thus, when I say that Edgar believes that water is wet, I am making the claim that he has in his belief box the Mentalese analogue of the English sentence "water is wet"; I am doing more than saying that he has a belief with the same content as that English sentence. This does seem a little implausible as it implies that folk psychology is committed to a specific theory concerning the nature of propositional attitudes, namely RTM. Surely most folk psychologists have never heard of, let alone endorsed, RTM.

Despite the fact that folk psychology does not individuate attitudes syntactically, it does distinguish between attitudes that agree in their truth conditional or referential content. For example, folk psychology does distinguish between believing that water is wet and believing that H<sub>2</sub>O is wet and most of those who would be happy to ascribe to Fang the belief that there is water in his bowl would resist the claim that Fang thereby believes that there is H<sub>2</sub>O in his bowl. This suggests that it is a fundamental assumption of folk psychology that there is more to a propositional attitude than a mode and a referential content and that extra component is not syntax. This makes it very tempting to conclude that folk psychology is committed to the existence of something like sense. If this conclusion were correct, then Fodor's view, as a claim about folk psychology, would be mistaken. Of course it would be open to him to argue that folk psychology does employ a notion of sense but that it is mistaken to do so. However, one might wonder how appealing such a view

would be to Fodor given his desire to vindicate folk psychology and make a respectable scientific psychology out of it.

(ii) A second problem with Fodor's rejection of sense is that he runs the real danger of running together two levels that he stresses are distinct, namely the intentional level and the computational level that implements it. As we have seen, he holds that intentional laws are computationally or syntactically implemented. That the relationship between the intentional and the computational is one of phenomena and laws at one level, being implemented by lower level phenomena and laws, is appealed to by Fodor in dealing with the accusation (most famously levelled by Stich (1983); see also Devitt (1991)) that the endorsement of the view that the mind is a syntax-driven machine implies that intentional and semantic properties have no proper place in psychological explanation. Such a defence of intentional psychology - and the associated account of the nature of the relationship between the intentional and the syntactic/computational, and their respective places in the scientific hierarchy - would appear to be inconsistent with the idea that intentional psychology must individuate partly in terms of syntax; for that idea entails that intentional psychology must appeal to the syntactic properties of mental representations in framing its laws and explanations.

It is clear that intentional psychology needs more than just mode and referential content. A psychology that recognised, appealed to, and individuated propositional attitudes in terms of just their mode and their reference would be descriptively and explanatorily inadequate. This is because beliefs can agree in their referential content yet diverge in their causal powers; for example *water* beliefs have different causal powers from *H<sub>2</sub>O* beliefs. Consequently, if one is going successfully to predict an individual's behaviour one needs to know more than the referential content of her belief. Similarly, if one is going to explain adequately an individual's behaviour one will need to specify more than the referential content of the belief appealed to.

As a consequence of the above described facts, the laws and generalisations employed and alluded to in the course of constructing intentional explanations and predictions will need to appeal to more than the referential content of propositional attitudes.

If the generalisations of intentional psychology are such that to satisfy the antecedent of such a generalisation all that is required of a subject is that she has an attitude with a certain mode and a certain referential content then they will be at worst flat false, and at best useless for the purposes of intentional prediction and explanation.

Hence, intentional psychology must recognise the existence of, appeal to, and individuate propositional attitudes in terms of, properties other than modes and referential contents. A psychology that ignored such differences as that between believing that water is wet and believing that H<sub>2</sub>O is wet would be descriptively and explanatorily inadequate.

(iii) A third problem with Fodor's rejection of sense has to do with behaviour. For very familiar reasons behaviour cannot be reduced to physical movement; type-identical behaviour can involve the making of very different physical movements, and identical physical movements can constitute divergent behaviour (Pylyshyn (1983)). Behaviour has semantic and intentional properties in terms of which we describe and individuate it. Some recent behaviour of Edgar includes hiding from Fang, ringing the emergency services, and asking for directions to the nearest hospital. Behaviour inherits its semantic and intentional properties from its mental causes, and from features of the wider context in which it takes place. For example, whilst riding a cycle Waldo sticks out his left arm and in so doing indicates to turn left. His behaviour is that of indicating to turn left partly in virtue of its intentional causes, (that is, the beliefs, desires, intentions etc. that figure in its etiology), and partly because of such facts about the wider context as that it is a convention in our society that one indicates to turn left on a cycle by sticking out one's left arm.

Now just as sentences that attribute propositional attitudes to an individual are opaque to the substitution of co-referential expressions, so are sentences that describe an individual's behaviour. The fact that it is true that an individual is searching for water doesn't imply that it is true that she is searching for H<sub>2</sub>O. There is a difference between water behaviour and H<sub>2</sub>O behaviour, just as there is a difference between gold behaviour and stuff with atomic number 79 behaviour, and so on. Ordinary every day folk recognise and honour such differences in describing and individuating the



behaviour of themselves and their fellows. Waldo is prospecting for gold in a respect in which he is not prospecting for stuff with atomic number 79; Fang is scratching around for water in a respect in which he is not scratching around for H<sub>2</sub>O; and Edgar is setting a trap to catch the thief of his prize delphiniums in a respect in which he is not setting a trap to catch Waldo despite the fact that Waldo is the culprit.

Thus, in describing and individuating behaviour, folk psychology appeals to, and must appeal to, properties of behaviour other than, so to speak, referential content. In this respect behaviours are just like propositional attitudes. How are we to account for this fact? What is the difference between water behaviour and H<sub>2</sub>O behaviour? It is clearly no good to argue that the difference is syntactic, as behaviour doesn't have syntactic properties; only certain kinds of symbols have syntactic properties. My thought is that the difference is semantic, something like a difference in sense. In saying this, the conclusion that I am driving at is that in abandoning sense Fodor has no way of accounting for such differences in behaviour, for here syntax will not do the job that sense is standardly invoked to do. A psychology based on the idea that there is nothing more to meaning than reference would therefore miss important distinctions between different types of behaviour; in other words, its taxonomy of behaviour would be far too coarse-grained.

It might be objected to this argument that one can account for the difference between water and H<sub>2</sub>O without invoking some second component of meaning or arguing that behaviour has syntactic properties. The idea is that we describe and individuate behaviour in terms of what the subject has in mind, in terms of how she sees or represents to herself that behaviour. This won't do for at least two reasons. Firstly, animals often behave without having anything in mind or without representing to themselves what they are doing. Is it really plausible to say that as a thirsty Fang scratches around for water he sees himself as scratching around for water? Secondly, often when people do have a conception of what they are doing they are mistaken so that the correct description of what they were doing wouldn't be a correct description of what they had in mind in so acting. Consider, for example, those lonely folks who go shopping everyday in a desperate search for companionship. As far as they are

concerned what they are doing is going shopping, yet what they are really doing is searching for companionship, despite how they represent their behaviour to themselves.

Thus, I conclude that Fodor has failed to overcome the need for a second component of meaning; his attempt to invoke syntax to do the job normally given to sense fails.<sup>141</sup> The upshot of this is that his theory of content is at best incomplete, as it naturalises only one component of meaning, namely, reference. His theory might tell us why, for example, the thought that I express with the English words "water is wet" is about water. But it doesn't tell us why it has the same content as the English sentence "water is wet" rather than that of "H<sub>2</sub>O is wet".

## 7.8 Conclusion

In this chapter I have presented a whole battery of objections to Fodor's naturalistic theory of content and therefore conclude that that theory fails. By way of recapitulation my central claims are as follows: (i) Fodor's theory does not apply to the simple non-logical symbols of natural language. Consequently, the theory is robbed of much of its motivation. Moreover, given the close relationship between language and thought, a failure to apply to natural language comes very close to implying a failure to apply to Mentalese. (ii) Fodor's theory requires that there are nonintentional causal laws relating the tokening of symbols to the instantiation of the properties that they express. Yet with respect to many of the symbols that express the contents of personal level propositional attitudes, there are no such laws; the only laws that govern the tokening of such symbols are inherently intentional. (iii) At the subpersonal level there are, arguably, symbols whose tokening is governed by the kinds of laws that Fodor requires. However, in these cases there is no

---

<sup>141</sup> It might be thought that one option open to Fodor is that of appealing to narrow content. The idea is that the difference between my water thoughts and my H<sub>2</sub>O thoughts is a difference in narrow content and that such differences are reflected in the syntactic properties of the sentences that we use to ascribe thoughts to ourselves and our fellows. However, there is no mileage in this idea as water thoughts have the same narrow content as their H<sub>2</sub>O thought analogues (as they instantiate the same function from contexts to referential contents).

asymmetric dependence. In general, where there are nonintentional laws governing the tokening of a symbol, there are no relationships of asymmetric dependence between those laws. And in those cases where there is something approaching asymmetric dependence (as in the case of the tokening of such natural language symbols as "horse") there are no nonintentional laws in operation. (iv) There are a number of problem cases; that is, cases where Fodor's theory implies that a symbol has a meaning at odds with its actual meaning. (v) Fodor's theory is essentially a theory of reference. This is a weakness if there is more to the content of our mental states than their reference or if scientific psychology needs to appeal to a further dimension of mental meaning. Fodor's attempt to invoke syntax to play the role of that extra dimension fails.

What are we to do in the light of the failure of Fodor's theory? What morals should be drawn from my reflections? We should certainly not abandon the naturalisation project. But an approach somewhat at odds with that adopted by Fodor is needed. Fodor attempted to construct - in one fell swoop and by relatively a prioristic means - an all encompassing theory of content. And he arrived at his conclusions on the basis of reflection on a small number of cases that, though familiar and salient from the folk psychological perspective, may not be all that typical. Given the differences between the various representational states that are attributed to us by scientific psychologists, there are good reasons to be sceptical of the existence of any general theory of content. In all probability, some of our representational states get their intentional and semantic properties determined in one way, and others in quite another way. Consequently, we should adopt a more piecemeal approach, concentrating our attention on only a small range of related intentional states at any one time. However, having said that, we should not focus our attention on too narrow a range of examples or we shall run the risk of making a mistake analogous to that made by Fodor as a result of his concentrating too much of his attention on thoughts about horses and thoughts about cows. Such a piecemeal approach cannot proceed in ignorance of, and indifference to, actual empirical research for we need to know what kinds of intentional states scientific psychology takes us to have, which specific intentional and semantic properties it attributes to those states, and

what the facts are surrounding their tokening. But there may well be an important role for a prioristic reasoning. If we are to have any realistic hope of uncovering the nonintentional and nonsemantic determinants of the semantic and intentional properties of a particular type of intentional state, we need to have a detailed understanding of the nature of the properties in question. In general, if one seeks to naturalise a property *P*, it is a good idea to have a detailed understanding of the nature of that property. It is not outlandish to say that such an understanding is precisely the sort that is arrived at by the successful construction by traditional philosophical means of a constitutive account of what it is to have the property in question. In other words, both traditional philosophical analysis and a consideration of actual scientific psychological theories and explanations, have an important place in any sensible approach to the naturalisation project. One might say that the project has a distinctly interdisciplinary air about it. This implies a criticism of Fodor's approach on two counts. First, he doesn't make enough of an attempt to understand and shed light on the nature of the properties that he seeks to naturalise. For example, he launches into a search for the nonsemantic and nonintentional determinants of the meaning of the Mentalese symbol that means *horse* without considering what is involved in having that meaning. Second, he nowhere considers or appeals to actual scientific psychological research and findings. I find this quite ironic given that Fodor spent much of his early and middle career telling us that there is no fundamental divide between philosophy and scientific psychology and that developments in the latter field can shed much light upon the traditional concerns of those engaged in the former.



# References

Antony, L. (1989). 'Anomalous Monism and the Problem of Explanatory Force'. *Philosophical Review* 98, pp. 153-187.

Baker, L.R. (1991). 'Has Content Been Naturalized?' In Loewer and Rey (1991).

Baker, L.R. (1995). *Explaining Attitudes: A Practical Approach to the Mind*. Cambridge: Cambridge University Press.

Bermudez, J.L. (1995). 'Nonconceptual Content: From Perceptual Experience to Subpersonal Computational States.' *Mind and Language*, 10, pp. 333-369.

Block, N. (1980) (ed.). *Readings in Philosophy of Psychology, Vol. i*. Cambridge, MA: Harvard University Press.

Block, N. (1981). (ed.) *Readings in Philosophy of Psychology, Vol. ii*. Cambridge, MA: Harvard University Press.

Block, N. (1986). 'Advertisement for a Semantics in Psychology'. *Midwest Studies in Philosophy*, 10, pp. 257-274. Reprinted in Stich and Warfield (1994).

Block, N. (1990a). 'The Computer Model of the Mind.' In D.N. Osherson and E.E. Smith (eds.) *An Invitation to Cognitive Science, iii: Thinking*. Cambridge, MA: MIT Press.

Block, N. (1990b). 'Can the Mind Change the World?' In Boolos (1990).

Block, N. (1991). 'What Narrow Content is Not'. In Loewer and Rey (1991).

- Boden, M. (ed.) (1990). *The Philosophy of Artificial Intelligence*. Oxford: Oxford University Press.
- Boghossian, P.A. (1989). 'The Rule Following Considerations'. *Mind* 98, pp. 507-549.
- Boghossian, P.A. (1991). 'Naturalizing Content'. In Loewer and Rey (1991).
- Boolos G. (ed.) (1990). *Meaning and Method: Essays in Honor of Hilary Putnam*. Cambridge: Cambridge University Press.
- Burge, T. (1979). 'Individualism and the Mental'. *Midwest Studies in Philosophy* 5, pp. 73-122.
- Burge, T. (1986a). 'Individualism and Psychology'. *Philosophical Review* 95, pp. 3-46.
- Burge, T. (1986b). 'Cartesian Error and the Objectivity of Perception'. In Pettit and McDowell (1986).
- Carruthers, P. (1996). *Language, Thought, and Consciousness*, Cambridge: Cambridge University Press.
- Chomsky, N. (1959). 'Review of Skinner's *Verbal Behaviour*'. *Language* 35, pp. 26-58.
- Churchland, P.M. (1979). *Scientific Realism and the Plasticity of Mind*. Cambridge: Cambridge University Press.
- Churchland, P.M. (1981). 'Eliminative Materialism and the Propositional Attitudes'. *Journal of Philosophy* 78, pp.67-90.
- Churchland, P.S. (1986). *Neurophilosophy*. Cambridge, MA: MIT Press.
- Cummins, R. (1983). *The Nature of Psychological Explanation*. Cambridge, MA: MIT Press.

- Cummins, R. (1989). *Meaning and Mental Representation*. Cambridge, MA: MIT Press.
- Davidson, D. (1963). 'Actions, Reasons, and Causes'. *Journal of Philosophy* 60. Reprinted in Davidson (1980).
- Davidson, D. (1970). 'Mental Events'. In L. Foster and J.W. Swanson (eds.) *Experience and Theory*. Amherst, MA: University of Massachusetts Press. Reprinted in Davidson (1980).
- Davidson, D. (1973). 'The Material Mind'. In P. Suppes, L. Henkin, G.C. Moisil, and A. Joja (eds.) *Proceedings of the Fourth International Congress for Logic, Methodology, and Philosophy of Science, Bucharest, 1971*. Reprinted in Davidson (1980).
- Davidson, D. (1974). 'Psychology as Philosophy'. In S.C. Brown (ed.) *Philosophy of Psychology*. London: Macmillan. Reprinted in Davidson (1980).
- Davidson, D. (1980). *Essays on Actions and Events*. Oxford: Oxford University Press.
- Davidson, D. (1987). 'Knowing One's Own Mind'. *Proceedings and Addresses of the American Philosophical Association*, pp. 441-458.
- Davidson, D. (1993). 'Thinking Causes'. In Heil and Mele (1993).
- Davies, M. (1986). 'Externality, Psychological Explanation, and Narrow Content'. *Proceedings of the Aristotelian Society Supplementary Volume* 60, pp. 263-283.
- Davies, M. (1989). 'Tacit Knowledge and Subdoxastic States.' In George (1989).
- Davies, M. (1991). 'Individualism and Perceptual Content'. *Mind* 100, pp. 461-484.

Davies, M. (1992). 'Perceptual Content and Local Supervenience'. *Proceedings of the Aristotelian Society*, 92, pp. 21-45.

Dennett, D.C. (1969). *Content and Consciousness*. London: Routledge.

Dennett, D.C. (1978a). *Brainstorms; Philosophical Essays on Mind and Psychology*. Montgomerly, VT: Bradford Books.

Dennett, D.C. (1978b). 'Artificial Intelligence as Philosophy and as Psychology.' In M. Ringle, ed. (1978). *Philosophical Perspectives on Artificial Intelligence*. New York: Humanities Press. (Reprinted in Dennett, 1978a).

Dennett, D.C. (1978c). 'A Cure for the Common Code?' In Dennett (1978a).

Dennett, D.C. (1981). 'Three Kinds of Intentional Psychology.' In R. Healy, (ed.) *Reduction, Time and Reality*. Cambridge: Cambridge University Press. (Reprinted in Dennett, 1987a)

Dennett, D.C. (1982a). 'Beyond Belief.' In Woodfield (1982). (Reprinted in Dennett, 1987a).

Dennett, D.C. (1982b). 'Styles of Mental Representation.' *Proceedings of the Aristotelian Society* LXXXIII, pp. 213-226. (Reprinted in Dennett, 1987a).

Dennett, D.C. (1987a). *The Intentional Stance*. Cambridge MA: MIT Press.

Dennett, D.C. (1987b). 'Evolution, Error and Intentionality'. In Y. Wilks and D. Partridge (eds.) *Source Book on the Foundations of Artificial Intelligence*. Cambridge: Cambridge University Press. Reprinted in Dennett (1987a).

Devitt, M. (1989). 'A Narrow Representational Theory of Mind'. In Silvers (1989).



Devitt, M. (1991). 'Why Fodor Can't Have it Both Ways'. In Loewer and Rey (1991).

Dretske, F.I. (1981). *Knowledge and the Flow of Information*. Cambridge, MA: MIT Press.

Dretske, F. I. (1986), 'Misrepresentation'. In R. Bogdan (ed.) *Belief*. Oxford: Oxford University Press.

Dretske, F.I. (1988). *Explaining Behaviour*. Cambridge, MA: MIT Press.

Dretske, F.I. (1990). 'Does Meaning Matter?'. In E. Villanueva (ed.) *Information, Semantics and Epistemology*. Oxford: Basil Blackwell.

Egan, F. (1991). 'Must Psychology be Individualistic?' *The Philosophical Review* 100, pp. 179-203.

Egan, F. (1992). 'Individualism, Computation and Perceptual Content'. *Mind* 101, pp. 443-459.

Egan, F. (1994). 'Individualism and Vision Theory'. *Analysis* 54, pp. 258-264.

Feigl, H. (1958). 'The "Mental" and the "Physical"'. In H. Feigl, M. Scriven and G. Maxwell (eds.) *Minnesota Studies in the Philosophy of Science*, vol. 2. Minneapolis: University of Minnesota Press.

Feyerabend, P. (1963). 'Materialism and the Mind Body Problem'. *Review of Metaphysics* 17, pp.49-67.

Field, H. (1978). 'Mental Representation'. *Erkenntnis*, 13, pp. 9-61. Reprinted in Block (1981).

Fodor, J.A. (1968a). 'The Appeal to Tacit Knowledge in Psychological Explanation'. *Journal of Philosophy* 65, pp.627-640. Reprinted in Fodor (1981a).

Fodor, J.A. (1968b). *Psychological Explanation*. New York: Random House.

Fodor, J.A. (1974). 'Special Sciences'. *Synthese* 28, pp. 97-115. Reprinted in Fodor (1981a).

Fodor, J.A. (1975). *The Language of Thought*. New York: Thomas Y. Crowell.

Fodor, J.A. (1978a). 'Tom Swift and his Procedural Grandmother'. *Cognition* 6, pp. 229-247.

Fodor, J.A. (1978b). 'Three Cheers for Propositional Attitudes'. In E. Cooper and E. Walker (eds.) *Sentence Processing*. Hillsdale, NJ: Erlbaum. Reprinted in Fodor (1981a).

Fodor, J.A. (1978c). 'Propositional Attitudes'. *The Monist* 61, pp.501-523. Reprinted in Fodor (1981).

Fodor, J.A. (1980). 'Methodological Solipsism Considered as a Research Strategy in Cognitive Psychology'. *Behavioural and Brain Sciences* 3, pp. 63-109. Reprinted in Fodor (1981a).

Fodor, J.A. (1981a). *RePresentations*. Cambridge, MA: MIT Press.

Fodor, J.A. (1981b). 'The Present Status of the Innateness Controversy'. In Fodor (1981a).

Fodor, J.A. (1983). *The Modularity of Mind*. Cambridge, MA: MIT Press.

Fodor, J.A. (1984). 'Semantics Wisconsin Style'. *Synthese* 59, pp. 231-250. Reprinted in Fodor (1990a).

Fodor, J.A. (1985). 'Fodor's Guide to Mental Representation'. *Mind* 94, pp. 66-100.

- Fodor, J.A. (1987). *Psychosemantics*. Cambridge, MA: MIT Press.
- Fodor, J.A. (1989a). 'Why Should the Mind be Modular?'. In George (1989). Reprinted in Fodor (1990a).
- Fodor, J.A. (1989b). 'Making Mind Matter More'. *Philosophical Topics* 67, pp. 59-79. Reprinted in Fodor (1990a).
- Fodor, J.A. (1989c). 'Substitution Arguments and the Individuation of Belief'. In G. Boolos (1989). Reprinted in Fodor (1990a).
- Fodor, J.A. (1990a). *A Theory of Content and Other Essays*. Cambridge, MA: MIT Press.
- Fodor, J.A. (1990b). 'A Theory of Content, I: The Problem'. In Fodor (1990a).
- Fodor, J.A. (1990c). 'A Theory of Content, II: The Theory'. In Fodor (1990a).
- Fodor, J.A. (1990d). 'Psychosemantics, or Where do Truth Conditions Come From?'. In Lycan (1990).
- Fodor, J.A. (1991a). 'A Modal Argument for Narrow Content'. *Journal of Philosophy* 88, pp. 5-26.
- Fodor, J.A. (1991b). 'Replies'. In Loewer and Rey (1991).
- Fodor, J.A. (1994). *The Elm and the Expert*. Cambridge, MA: MIT Press.
- Fodor, J.A., Bever, T, and Garrett, M. (1974). *The Psychology of Language*. New York: McGraw Hill.
- Fodor, J.A. and Chihara, C. (1965). 'Operationalism and Ordinary Language'. *American Philosophical Quarterly* 2, pp. 281-295. Reprinted in Fodor (1981a).

Fodor, J.A. and LePore, E. (1992). *Holism: A Shopper's Guide*. Oxford: Basil Blackwell.

Fodor, J.A. and Pylyshyn, Z. (1988). 'Connectionism and Cognitive Architecture: A Critical Analysis'. *Cognition* 28, pp. 3-71.

George, A. (ed. ) (1989). *Reflections on Chomsky*. Oxford: Basil Blackwell.

Gibson, J.J. (1979). *The Ecological Approach to Visual Perception*. Boston: Houghton Mifflin.

Grimm, R. and Merrill, D. (eds.) (1988). *Contents of Thought*. Tucson: The University of Arizona Press.

Harman, G. (1972). *Thought*. Princeton, NJ: Princeton University Press.

Harman, G. (1982). 'Conceptual Role Semantics', *Notre Dame Journal of Formal Logic* 23, pp. 242-256.

Haugeland, J. (1981). *Mind Design*. Cambridge, MA: MIT Press.

Haugeland, J. (1985). *Artificial Intelligence: The Very Idea*. Cambridge, MA: MIT Press.

Hempel, C. (1980). 'The Logical Analysis of Psychology'. In N. Block (1980).

Hempel, C. (1965). *Aspects of Scientific Explanation*. New York: Free Press.

Heil, J. and Mele, A. (eds.) (1993). *Mental Causation*. Oxford: Oxford University Press.

Horgan, T. (1994). 'Computation and Mental Representation'. In Stich and Warfield (1994).

- Horgan T. and Woodward J. (1985). 'Folk Psychology is Here to Stay.' *The Philosophical Review* 94, pp. 197-226. Reprinted in Lycan (1990).
- Jackson, F. and Pettit, P. (1988). 'Functionalism and Broad Content'. *Mind* 97, pp. 381-400.
- Johnson-Laird, P. (1988). *The Computer and the Mind: An Introduction to Cognitive Science*. London: Fontana.
- Kim, J. (1984). 'Concepts of Supervenience'. *Philosophy and Phenomenological Research*, 65, pp. 153-176. Reprinted in Kim (1993).
- Kim, J. (1980). 'Supervenience as a Philosophical Concept'. *Metaphilosophy*, 21, pp. 1-27. Reprinted in Kim (1993).
- Kim, J. (1993). *Supervenience and Mind*. Cambridge: Cambridge University Press.
- Kim, J. (1996). *Philosophy of Mind*. Oxford: Westview Press.
- Kripke, S. (1980). *Naming and Necessity*. Oxford: Basil Blackwell.
- Kripke, S. (1982). *Wittgenstein: On Rules and Private Language*. Oxford: Basil Blackwell.
- Lewis, D. (1973). *Counterfactuals*. Cambridge, MA: Harvard University Press.
- Loar, B. (1981). *Mind and Meaning*. Cambridge: Cambridge University Press.
- Loar, B. (1988). 'Social Content and Psychological Content'. In Grimm and Merrill (1988).
- Loar, B. (1991). 'Can We Explain Intentionality?'. In Loewer and Rey (1991).



Loewer, B. and Rey, G. (eds.) (1991). *Meaning in Mind: Fodor and his Critics*. Oxford: Basil Blackwell.

Lycan, W.G. (1981). 'Toward a Homuncular Theory of Believing.' *Cognition and Brain Theory* 4, pp. 139-159.

Lycan, W.G. (ed.) (1990). *Mind and Cognition: A Reader*. Oxford: Basil Blackwell.

Marr, D. (1977). 'Artificial Intelligence - A Personal View'. *Artificial Intelligence* 9, pp. 37-48. Reprinted in Boden (1990) and Haugeland (1981).

Marr, D. (1982). *Vision*. San Fransisco: Freeman.

McCulloch, G. (1985). *The Mind and its World*. London: Routledge.

McDowell, J. (1986). 'Singular Thought and the Extent of Inner Space'. In Pettit and McDowell (1986).

McDowell, J (1994a). *Mind and World*. Cambridge, MA: Harvard University Press.

McDowell, J. (1994b). 'The Content of Perceptual Experience.' *The Philosophical Quarterly* 44, pp. 190- 205.

McGinn, C. (1984). *Wittgenstein on Meaning*. Oxford: Basil Blackwell.

McGinn, C. (1989). *Mental Content*. Oxford: Basil Blackwell.

Millikan, R.G. (1984). *Language, Thought and Other Biological Categories*. Cambridge, MA: MIT Press.

Millikan, R.G. 'Thoughts Without Laws: Cognitive Science Without Content'. *Philosophical Review* 95, pp. 47-80.

Millikan, R.G. (1989). 'Biosemantics'. *Journal of Philosophy* 86, pp. 281-297. Reprinted in Stich and Warfield (1994).

- Minsky, M. (1981). 'K-Lines: A Theory of Memory.' In Norman, D. (ed.), *Perspectives on Cognitive Science*. Norwood, NJ: Ablex.
- Newell, A. and Simon, H.A. (1976). 'Computer Science as Empirical Enquiry.' *Communications of the Association for Computing Machinery* 19, pp. 113-126. (Reprinted in Haugeland, 1981).
- Newell, A. (1980). 'Physical Symbol Systems.' *Cognitive Science* 4, pp. 87-127.
- Papineau, D. (1987). *Reality and Representation*. Oxford: Basil Blackwell.
- Papineau, D. (1993). *Philosophical Naturalism*. Oxford: Basil Blackwell.
- Patterson, S. (1996). 'Success-orientation and Individualism in Marr's Theory of Vision'. In K. Akins (ed.) *Perception*. Oxford: Oxford University Press.
- Peacocke, C. (1992). *A Study of Concepts*. Cambridge, MA: MIT Press.
- Peacocke, C. (1993). 'Externalist Explanation.' *Proceedings of Aristotelian Society*, 93, pp. 203-230.
- Peacocke, C. (1994). 'Content, Computation and Externalism.' *Mind and Language* 9, pp. 303-335.
- Pettit, P. and McDowell, J. (eds.) (1986). *Subject, Thought and Context*. Oxford: Oxford University Press.
- Place, U.T. (1956). 'Is Consciousness a Brain Process'. *British Journal of Psychology* 47, pp. 44-50.
- Putnam, H. (1975a). *Mind, Language and Reality: Philosophical Papers Volume 2*. Cambridge: Cambridge University Press.

Putnam, H. (1975b). 'The Meaning of "Meaning"'. In K. Gunderson (ed.), *Minnesota Studies in the Philosophy of Science*, Vol. 7. Minneapolis: University of Minnesota Press. Reprinted in Putnam (1975a).

Pylyshyn, Z. (1984). *Computation and Cognition*. Cambridge, MA: MIT Press.

Quine, W.V.O. (1953). *From a Logical Point of View*. Cambridge, MA: Harvard University Press.

Quine, W.V.O. (1951). 'Two Dogmas of Empiricism', *Philosophical Review* 60, pp. 20-43 Reprinted in Quine (1953).

Quine, W.V.O. (1960). *Word and Object*. Cambridge, MA: MIT Press.

Rumelhart, D.E., McClelland, J.E. and the PDP Research Group (1986). *Parallel Distributed Processing: Explorations in the Microstructure of Cognition, Volume 1 Foundations*. Cambridge, MA: MIT Press.

Rorty, R. (1965). 'Mind-Body Identity, Privacy, and Categories'. *Review of Metaphysics* 19, pp. 24-54.

Rosch, E. (1973). 'On the Internal Structure of Perceptual and Semantic Categories'. In T.E. Moore (ed.) *Cognitive Development and the Acquisition of Language*. New York: Academic Press.

Rosch, E. (1975). 'Cognitive Representations and Semantic Categories'. *Journal of Experimental Psychology: General*, pp. 192-233.

Rosch, E. (1978). 'Principles of Categorization'. In E. Rosch and B. Lloyd (eds.), *Cognition and Categorization*. Hillsdale, NJ: Erlbaum.

Ryle, G. (1949). *The Concept of Mind*. Harmondsworth: Peregrine.

Searle, J. (1992). *The Rediscovery of Mind*. Cambridge, MA: MIT Press.

Segal, G. (1989). 'Seeing What Is Not There'. *Philosophical Review* 98, pp. 189-214.

Segal, G. (1991). 'Defence of a Reasonable Individualism'. *Mind* 100, pp. 485-494.

Segal, G. and Sober, E. (1991). 'The Causal Efficacy of Content'. *Philosophical Studies* 63. pp. 1-30.

Silvers, S. (ed.) (1989). *Re-Representations: Readings in the Philosophy of Mental Representation*. Dordrecht: Kluwer Academic Publishers.

Skinner, B.F. (1957). *Verbal Behaviour*. New York: Appleton-Century-Crofts.

Smart, J.C.C. (1959). 'Sensations and Brain Processes'. *Philosophical Review* 68, pp. 141-156.

Stalnaker, R. (1984). *Inquiry*. Cambridge, MA: MIT Press.

Stalnaker, R. (1989). 'On What's in the Head'. In J.E. Tomberlin (ed.) *Philosophical Perspectives, Volume 3, Philosophy of Mind and Action Theory*.

Stampe, D. (1977). 'Towards a Causal Theory of Linguistic Representation'. *Midwest Studies in Philosophy* 2, pp. 42-63.

Stich, S.P. (1978a). 'Autonomous Psychology and the Belief-Desire Thesis.' *The Monist* 61, pp. 573-591.

Stich, S.P. (1978b). 'Beliefs and Subdoxastic States.' *Philosophy of Science* 45, pp. 499-518.

Stich, S.P. (1983). *From Folk Psychology to Cognitive Science*. Cambridge, MA: MIT Press.

Stich, S.P. (1992). 'What is a Theory of Mental Representation?' *Mind* 101, pp. 243-263. Reprinted in Stich and Warfield.

Stich, S.P. and Warfield, T. (eds.) (1994). *Mental Representation*. Oxford: Basil Blackwell.

Stillings et. al. (1987). *Cognitive Science: An Introduction*. Cambridge, MA: MIT Press.

Tye, M. (1992). 'Naturalism and the Mental'. *Mind*, 101, pp. 421-441.

van Gulick, R. (1989). 'Metaphysical Arguments for Internalism and Why They Don't Work'. In Silvers (1989).

Wittgenstein, L. (1953). *Philosophical Investigations*, trans. G.E.M. Anscombe. Oxford: Basil Blackwell.

Wittgenstein, L. (1958). *The Blue and the Brown Books*. Oxford: Basil Blackwell.

Woodfield, A. (ed.) (1982). *Thought and Object*. Oxford: Clarendon Press.

Wright, C. (1989). 'Wittgenstein's Rule-following Considerations and the Central Project of Theoretical Linguistics'. In George (1989).

Yablo, S. (1992). 'Mental Causation'. *Philosophical Review* 101. pp. 245-280.

Yuille, A.L. and Ullman, S. (1990). 'Computational Theories of Low-Level Vision'. In D. Osherson, S. Kosslyn, and J. Hollerbach (eds.) *An Invitation to Cognitive Science, Vol 3: Visual Cognition and Action*.